Building a text corpus for representing the variety of medical language

Benoît Habert[a], Natalia Grabar[b], Pierre Jacquemart[b], Pierre Zweigenbaum[b]

[a] LIMSI-CNRS & Université Paris 10

[b] DIAM - Service d'Informatique Médicale/DSI, Assistance Publique – Hôpitaux de Paris & Département de Biomathématiques, Université Paris 6

Abstract

The representation of specialized domains in reference corpora does not always cater for the internal diversity of genres. Similarly, most sublanguage studies have focussed on domain specialization, largely leaving genre an implicit choice that received less individual attention. Specialized domains, though, display a large palette of text genres. Medicine is a case in point, and has been the subject of much work in Natural Language Processing. We therefore endeavored to build a large corpus of medical texts with a representation of the main genres found in that domain. We propose a framework for designing such a corpus: an inventory of the main genres of the domain, a set of descriptive dimensions and a standardized encoding of both meta-information (implementing these dimensions) and content. We present a proof of concept demonstrator encoding an initial corpus of text samples according to these principles.

Keywords:

Specialized language, genres, medicine, French, natural language processing, TEI, CES, XML, XSL.

1 Introduction

The representation of specialized domains in reference corpora does not always cater for the internal diversity of genres. This is the case for instance in Brown and LOB, where the domain is the entry key – the small size of these corpora also accounts for this. Similarly, domain specialization seems to have been the focal criterion considered in all sublanguage studies performed in the eighties (*e.g.*, (Grishman & Kittredge, 1986)), whereas genre was largely an implicit choice that received less individual attention.

Specialized domains, though, display a large palette of text genres. This is the case of medicine. Medical narratives, including discharge summaries and imaging reports, have been the most studied type of text (Sager *et al.*, 1987; Friedman, 1997; Rassinoux, 1994; Zweigenbaum & Consortium MENELAS, 1994). Short problem descriptions, such as signs, symptoms or diseases, have been the subject of much attention too, in relation to standardized vocabularies (Tuttle *et al.*, 1998). Some authors have also examined abstracts of scientific literature (Grefenstette, 1994). And indeed, web pages are today the most easily available source of medical documents. These documents vary both in form and in content; it has even been showed that within a single document, subparts can consistently display very different language styles (Biber & Finegan, 1994). The natural language processing (NLP) tools that have been tailored for one document type may therefore be difficult to apply to another genre (Friedman, 1997)[1]. All these genres are found in the same overall domain: medicine. A large palette of genres can also be found within each medical specialty: domain and genre are clearly distinct factors of textual variety.

This diversity has consequences for the design and development, or simply for the use, of natural language processing tools for medical information processing. Without better informed knowledge about the differential performance of natural language processing tools on a variety of medical text types, it will be difficult to control the extension of their application to different medical documents. We propose here to provide a basis for such informed assessment: the construction of a corpus of medical texts. We address this task for French language texts, but we believe the same reasoning and methods and part of the results are applicable to other languages too.

This text corpus must be useful for testing or training NLP tools [2]. It must provide a variety of medical texts: diversity must be obtained in addition to mere volume, since our specific aim is to represent the many different facets of medical language. We need to characterize this diversity by describing it along

[1] The precision of French taggers evaluated within the framework of GRACE (Adda *et al.*, 1999), measured in relation to a manually tagged reference corpus, similarly shows significant variations depending on the part of the corpus under examination (Illouz, 1999). This corpus containing 100,000 words has been compiled from extracts from *Le Monde* (2 extracts), and from literary texts: memoirs (2 extracts), novels (6 extracts), essays (2 extracts). Thus an extract from memoirs results in important variations, positive and negative, among the taggers.

[2] Taggers, shallow parsers which are able to build syntactic representations for any kind of text, named entities recognizers, etc.

appropriate dimensions: origin, genre, domain, etc. These dimensions have to be documented precisely for each text. This documentation must be encoded formally, as meta-information included with each document, so that sub-corpora can be extracted as needed to study relevant families of document types. Finally, text contents must also be encoded in a uniform way, independently of the many formats documents were written in originally.

We present here a framework for designing a corpus of medical texts representing genre variety: a set of genres and descriptive dimensions, inspired in part from previous relevant literature, a standardized encoding of both meta-information (implementing these dimensions) and content, using the TEI XML Corpus Encoding Standard (Ide *et al*., 1996), and an initial set of text samples encoded according to these principles. This work takes place in the context of a larger corpus collection initiative, project CLEF (www.biomath.jussieu.fr/CLEF/), whose goal is to build a large, diversified corpus of French texts and to distribute it widely to researchers.

After a brief review of related work (section 2), we explain in turn each of the main phases of the design of our corpus: (*i*) assessing document diversity, choosing dimensions to characterize this diversity, and implementing them in a standard XML DTD (section 3); (*ii*) selecting the main classes of documents we want to represent and documenting them with these dimensions, then populating the corpus with texts (section 4). We also explain how sub-corpora can be extracted from the corpus (section 4.3).

2 Taking into account corpus state of the art

The evolution of corpus techniques and standards in the past ten years makes it difficult to *reuse* existing (medical) corpora.

The development of standards for the encoding of textual documents has been the subject of past initiatives in many domains (electronic publishing, aeronautics, etc.), using the SGML formalism, and now its XML subset. The Text Encoding Initiative was a major international effort to design an encoding standard for scholarly texts in the humanities and social sciences, including linguistics and natural language processing. It produced document type definitions (DTDs) which have been complemented with a Corpus Encoding Standard (CES) (Ide *et al*., 1996). The CES DTD is therefore the natural format for encoding a corpus that is targeted at NLP tools.

Beyond bibliographic description, descriptive dimensions for characterizing text corpora have been proposed by Sinclair (Sinclair, 1996) and Biber (Biber, 1994) among others. A related strand of work is that around the standardization of meta-information for documenting web pages (Dublin Core Metadata Initiative, 1999); but this covers more limited information than that we shall need. In the medical informatics domain, the standardization efforts of bodies such as HL7 (Dolin *et al*., 2000) and CEN (Rossi Mori & Consorti, 1999) focus on clinical documents for information interchange: both their aim and coverage are different from ours.

The available medical corpora we are aware of do not match the criteria underlying current standards. Firstly, medical textbooks and scientific literature have been collected in project LECTICIEL (Lehmann *et al.*, 1995) for French for Special Purposes learning. A set of software tools were available to study the various parameters of documents from the corpus (lexical choices, grammatical connectors, text organization: titles, parts...). Users could add new texts to the database and compare them with the existing sub-corpora. The encoding standard however is an obsolete one and the resulting corpus far too small by our current expectations. Secondly, one medical corpus was specifically built for the purpose of linguistic study: MEDICOR (Vihla, 1998). Although its focus is on published texts (articles and books), with no clinical documents, it is an example of the kind of direction that we wish to take. Unfortunately the initial version of the corpus provides limited documentation about the features of each document (intended audience, genre and writer qualification), which is planned to be extended. Thirdly, even though very large collections of medical texts indeed exist within hospital information systems – the DIOGENE system being among the earliest ones (Scherrer *et al.*, 1996) – the issue here is that of privacy and therefore anonymization, to which we return below.

3 Identifying and representing diversity dimensions

3.1 Assessing variety dimensions

In our opinion, a corpus can only represent some limited subsets of the language, and not the whole of it. No corpus can contain *every* type of language use[3]. It is even true in specialized domains such as medicine or computer science. As a matter of fact, studies of *sublanguages* favored until recently very few types of textual documents (see above), "hiding" the variety of registers within each of these sublanguages.

In order to gather a corpus, one must explicitly choose the language use(s) (s)he wants to focus on. One must identify the main underlying dimensions of diversity which are responsible for the major contrasts within the linguistic area (s)he wants to analyze. The variety factors are twofold: external and internal. External variety refers to the whole range of parameter settings involved in the creation of a document: document producer(s), document user(s), context of production or usage, mode of publication, etc. This issue is thoroughly addressed in (Sinclair, 1996) and (Biber, 1994). It is rather straightforward to describe documents according to these lines. However, internal variety must as well be taken into account. It follows from the range of registers corresponding to the main communicative tasks of the linguistic community. Indeed informants in a specific domain such as medicine have intuitions about the major relevant registers for the domain, even if they do have difficulties in establishing clear-cut borderlines. (Wierzbicka, 1985) relies on folk names of genres (*to give a talk / a paper / an address / a lecture / a speech*) as an important source of insight inside communicative characteristics of a given community. It has even been shown as well that, while there is no well-established genre palette for Internet materials, it is nevertheless possible, through interviewing users of Internet (students and teaching staff in computer science), to define genres that are both reasonably consistent with what users expect and conveniently computable using measures of stylistic variation (Dewe *et al*., 1998)[4]. This is why the very first step consists in asking people from the domain the main communicative routines or "speech acts" they identify. We started thus from a series of prototypic contexts, and listed the types of texts related to these starting points: medical doctor (in hospital or in town), medical student, patient (consumer); patient care, research; published and unpublished documents.

It is now possible to restate more precisely what we mean by variety : a domain corpus should represent the *main* communicative acts of the domain and their parameter settings. This analysis leads us to slightly change Sinclair's definition of a corpus (Sinclair, 1996, p. 4) ("a corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language"): the criteria need to be situational and sociological as well (Habert, 2000). The definition of the dominant communicative routines of the domain and of their precise situational parameters is a pre-requisite.

3.2 Genres of medical texts

Trying to compile a complete list of the types of medical textual documents is probably a never ending task, since new situations may lead to the creation of new document types, and since a finer-grain examination could always reveal finer distinctions. Our aim here is rather to identify the main kinds of medical texts that can be found in computerized form, and to characterize each of them by specifying values for a fixed set of orthogonal dimensions related to the external and internal factors of diversity.

We considered four main contexts of production or use of medical documents (table 1).

In the context of care, medical professionals produce information about a patient. In a hospital, this information is registered in the form of reports. Letters are a different series of genres where the receiver is a more targeted health care professional.

In a university context, teaching involves material produced by faculty members (lecture notes as well as test questions for examinations) and by students (student notes).

Dissemination of knowledge much resembles what can be found in other scientific disciplines. Medical professionals read and, for some of them, write articles in various sorts of journals and conference proceedings. Medical professionals or scientific reporters also write articles for the general public in the

[3]For instance, is it possible to find real-life, that is large-scale, samples of prayers or love small-talk in any existing corpus?

[4]The ten following genres were distingued : Informal, Private (Personal home pages); Public, commercial (Home pages for the general public); Interactive pages (Pages with feed-back: customer dialogue; searchable indexes); Journalistic materials (News, editorials, reviews, popular reporting, e-zines); Reports (Scientific, legal, and public materials; formal text); Other running text; FAQs; Link Collections; Other listings and tables; Discussions; Contributions to discussions; Usenet News material; Error Messages.

general press or in more specialized magazines. They also report on (often yet unpublished) conference lectures. Students also produce memos and reports, the most prominent of which is the doctoral dissertation. Direct computer mediated discussion and exchange of information takes place in specialized newsgroups and electronic lists, which are as pervasive in the medical area as in other technico-scientific domains.

Production of information on a patient	*Reports*	Discharge report Surgery report *Examination report* Endoscopy ; EKG; EEG; Anatomopathology; Imaging ; Functional
	Letters	Request for advice; Referral; Discharge letter; Additional prescription
Teaching	*Course material (professorial)*	Lecture notes; Test questions
	Course material (student)	Student notes
Dissemination of knowledge	*Periodicals*	Newspaper ; Magazine ; Bulletin ; Journal
	Articles	Generalist press article; Scientific article; Article abstract; Conference report
	Student memo	Dissertation
	Electronic discussion	News group ; Electronic list
Knowledge resources	*Reference knowledge*	Book; Encyclopedia; Dictionary; Monograph
	Guidelines	General French medical guidelines; Consensus conference; Recommendation; Protocol
	Official	Bulletin officiel; Code of deontology; Convention; Informed consent
	Coding systems	Terminology (Nomenclature, Thesaurus, Classification)

Table 1: A (non-exhaustive) list of genres of medical documents.

Beyond the previous items, different kinds of knowledge resources are used in the medical practice. Stable reference knowledge is found in dictionaries and encyclopedias or in monographs (*e.g.*, all that needs to be known about a given drug). More operative knowledge takes the form of guidelines and protocols which often constitute rules that medical practitioners must follow. The "Références médicales opposables" (RMO, http://www.upml.fr/rmo/, translated here as "general French medical guidelines") are national rules that state which medical acts should be avoided in certain situations. A "consensus conference" is a conference where an authoritative group of physicians agree on a statement, *e.g.*, about the best treatment for a disease (*e.g.*, www.chu-rouen.fr/ssf/recomfr.html). Protocols are precise plans of diagnosis or treatment for specific diseases, especially in oncology. Official documents regulate the legal or contractual aspects of medical practice: the official bulletin is a legal publication of the French government; the Code of deontology is a regulation of the medical profession; the "Convention" is an agreement between the medical profession and the government; and the "Informed consent" warns a patient about the potential risks and benefits related to his or her treatment. Finally, coding systems organize different types of terminologies used for the normalized description of medical information.

3.3 Dimensions

These document types are difficult to classify into non-overlapping groups. Therefore modeling the corpus with descriptive dimensions is all the more useful. We studied three sets of dimensions proposed in the literature (Sinclair, 1996; Biber, 1994; Dublin Core Metadata Initiative, 1999). Most of them are attributes useful in medical text genres, and were kept in our final selection (table 2). The objective of these attributes is to characterize the different types of texts and each of their instances. Here again, the set presented here is liable to revision as more documents are added to the corpus.

We divide these dimensions in three groups. Bibliographic dimensions are the traditional features that characterize the origin of the document, including its links to a larger document set (*e.g.*, article in a journal or chapter in a book) and its status as a partial or full document. Non-textual parts are removed from our documents, but their descriptions can be included.

External dimension: bibliographic references		
Origin :	*Title :* *Author, Coauthor, Translator, Contributor, Editor, Text creator:* *Date of creation, of translation, etc.:* *Identifier:*	
Localization:	Page, etc.	
Link:	None; To a series; To another text	
Extract:	Full; Article; Chapter; Paragraph	
Description of embedded non-textual data:	e.g., Radiograph, Photograph, Table	
External dimension: context of production and reception		
Mode of production:	Typed; Dictated; Manuscript	
Mode of transmission:	Oral; Electronic; Printed	
Software format:	Plain text, html, etc.	
Producer	*Plurality:*	Individual; Association; Company; Institution
	Function:	Scientist, Medical professional, Scientific reporter, Student, Terminologist, Patient, etc.
Receiver	*Plurality*:	Unique, Multiple
	Presence:	Present; Absent
	Profile:	Medical professional; Non-medical professional
Objective:	Record; Describe; Inform; Explain; Discuss; Persuade; Recommend; Teach; Order	
Publication status:	Published; Unpublished	
Frequency of publication:	Periodical; Punctual	
Coverage:	Local; National; International	
Rights:	Statement of ownership, Usage restrictions	
More internal dimensions		
Language:	French	
Size of text:	In words, bytes, etc.	
Level of style:	Low; Medium; High	
Quality of presentation:	Raw; Revised; Advanced	
Interaction with public:	Distant; Neutral; Close	
Personalization:	Personalize; Impersonal	
Factuality:	Informative factual; Intermediate; Imaginary	
Technicity:	Non-technical; Intermediate; Specialized	

Table 2: Dimensions.

The second group contains external dimensions that characterize the context of production or reception of the documents. The mode of production corresponds to the original authoring of the text, the mode of transmission to the form it had before inclusion in the corpus. Only the mode of transmission is considered in (Sinclair, 1996) and (Biber, 1994). We find it useful to make a difference between the successive forms of a document during its life cycle. The software format applies to electronic texts and helps to document conversion work. The producer's profile includes his or her "function". This aspect is not mentioned in (Sinclair, 1996), although it covers the profile of the receiver ("audience constituency"). The only

distinction in the receiver's profile that seems relevant up to now here is whether s/he is a medical professional or not. The "objectives" merge those of (Biber, 1994) and of (Sinclair, 1996). The publication status corresponds to a usual distinction. The frequency of publication was introduced as a general attribute to help differentiate periodicals from non-periodicals. "Coverage" comes from the Dublin Core (Dublin Core Metadata Initiative, 1999): it is meant to describe "the extent or scope of the content of the resource". Typically this encompasses the spatial and temporal validity of the text: *e.g.*, the national applicability of a law or the temporal validity of a terminology. The "rights" attribute, also from the Dublin Core, is useful to inform the corpus user about the allowed utilization of the corpus.

The last group contains more internal dimensions which can generally be detected from the text itself. "Language" is mentioned in (Dublin Core Metadata Initiative, 1999). We introduced "size" to control sampling policy over the corpus. The "level of style" is a usual dimension, as well as the "quality of presentation" (Sinclair, 1996). Some of the remaining dimensions may be related to external dimensions (*e.g.*, "Interaction with public" usually depends on the producer's and receiver's profiles), but we consider that they are more reflected and verifiable in text contents.

3.4 Domains

Medicine has numerous specialized subfields, each of which entertains professional, teaching and scientific activities, with its own societies, journals and conferences. Several lists of medical specialties can be found, among which the official list of health care professions ("Nomenclatures des professions de santé"), that of the US National Library of Medicine's Medical Subject headings thesaurus (MeSH, www.nlm.nih.gov/mesh/meshhome.html) and that of the CISMeF internet directory of French medical resources (www.cismef.org, (Darmoni *et al.*, 2000)). We relied on the latter, which synthesizes the first two. We made some of the categories slightly more specific by ungrouping some clusters (*e.g.*, "Angéiologie & cardiologie" separated into "Angéiologie" and "cardiologie"); we also removed or specialized a few themes that seemed too far fetched (*e.g.*, "Anthropology, Education, Sociology and Social Phenomena", reduced to "Education").

3.5 Exploiting diversity dimensions: corpus and document headers

A corpus without documentation is a (possibly huge) bag of "dead words". For that very reason, within the TEI standardization group, much attention has been devoted to the definition of *headers* (Giordano, 1995). A header is a normalized way of documenting electronic texts. The corpus header caters for the documentation for the corpus as a whole, whereas each document header contains the meta-information for its text.

Each document in a corpus has a header. This header describes the electronic text and its source – `fileDesc` or file description in figure 1– (bibliographic information, when available), it gives the encoding choices for the text – `encodingDesc` or encoding description – (editorial rationales, sampling policy...), non-bibliographical information that characterize the text, and a history of updates and changes (`revisionDesc` – revision description). In the non-bibliographical part of the header (`profileDesc` – profile description), the text is "tagged" according to one or more standard classification schemes, which can mix both free indexes and controlled ones (such as standard subject thesauri in the relevant field).

These classification schemes are thoroughly described in the *corpus header*. It is then possible to extract sub-corpora following arbitrarily complex constraints stated in these classification schemes. For instance, the interface to the BNC relies on such an approach (Dunlop, 1995) and permits to restrict queries to sub-corpora (spoken *vs* written language / publication date / domain / fiction *vs* non-fiction ...and any combination of these dimensions).

We followed the TEI proposals and more precisely the standard TEI XML CES model (Ide *et al.*, 1996). We could find a mapping into the CES header for each dimension of our model, and therefore implemented it in the CES framework. Generally, bibliographic dimensions (and document size) fit into the `fileDesc`; the definition of the other external and internal dimensions is located in the `encodingDesc` section of the corpus header, and each corpus document refers to it in its `profileDesc` section.

An added advantage is that the CES model provides some additional documentation dimensions, *e.g.*, information about the corpus construction process (text conversion, normalization, annotation, etc.). The implemented corpus is a collection of texts, each with its own meta-information: an instantiation of the above dimensions. On top of these texts, it provides documentation on itself: on the one hand, bibliographic

information of the same kind as its component texts; on the other hand, meta-information about the documents it contains. The latter comprises the definition of the descriptive dimensions along which each of its documents is described.

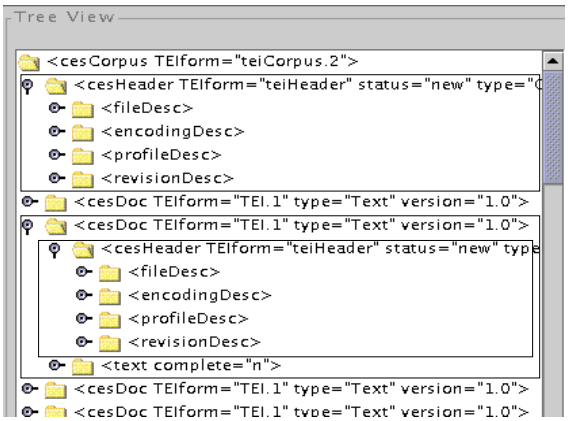

Figure 1: Overall corpus form: corpus header (<*cesHeader*>: upper rectangle) then documents (<cesDoc>), each containing a document header (<cesHeader>: lower, inner rectangle) and the actual <text>.

4 Building and exploiting the corpus

4.1 Giving a shape to the corpus: document sampling

Several parameters influence the overall contents of the corpus: we focus here on the types and sizes of documents that it will include. There is debate in the corpus linguistics community as to whether a corpus should consist of text extracts of constant size, as has been the case of many pioneering corpora, or of complete documents. The drawback with extracts is that textual phenomena with a larger span may not be studied on such samples. The overall strategy of project CLEF is therefore to opt for full documents as much as possible. In some cases however, text samples may be easier to obtain: it may be more acceptable for a publisher, because of property rights, to give away extracts rather than full books or journals. We plan to be pragmatic about this issue.

To initiate the construction of our corpus, we selected an initial subset of text types as target population for the corpus. As explained above, we tried to represent the main communicative acts of the domain. The main text types we aim to represent initially include types from all the groups of genres listed above: hospital reports, letters (discharge), teaching material (tutorials), publications (books chapters, journal articles, dissertations), guidelines (recommendations) and official documents (code of deontology). This will be achieved progressively; the current status is that of a proof of concept, which we describe below. We cautiously avoided to over-represent web documents, which could bias corpus balance because of their immediate ease of obtention. An additional interesting family of genres would be transcribed speech; but the cost of transcription is too high for this to be feasible.

A generic documentation for each text type was prepared. The rationale for implementation is then to encode a document header template for each text type: this template contains the prototypical information for texts of this type. This factorizes documentation work, so that the remaining work needed to derive a suitable document header for an individual text is kept to a minimum. Document templates were implemented for the text types included so far in the corpus.

4.2 Populating the corpus with document instances

The addition of documents to the corpus comprises several steps. The documents must first be obtained. This raises issues of property. A standard contract has been established for the project with the help of the European Language Resources Agency (ELRA), by which document providers agree with the distribution of the texts for research purposes. For texts that describe patient data, a second issue is that of

privacy. We consulted the French National Council for Informatics and Liberties (CNIL). They accepted that such texts be included provided that all proper names (persons and locations) and dates be masked.

The contents of each document are then converted from their original form (HTML, Word) to XML format. Minimal structural markup is added: that corresponding to the TEI CES level 1 DTD. This includes paragraphs (<*p*>; this is marked automatically) and optionally sections (<*div*>). The document header template for the appropriate document type is then instantiated. For series of similar samples (*e.g.*, a series of discharge summaries), most of this instantiation can be performed automatically.

Figure 2 shows a slice of the implemented corpus.

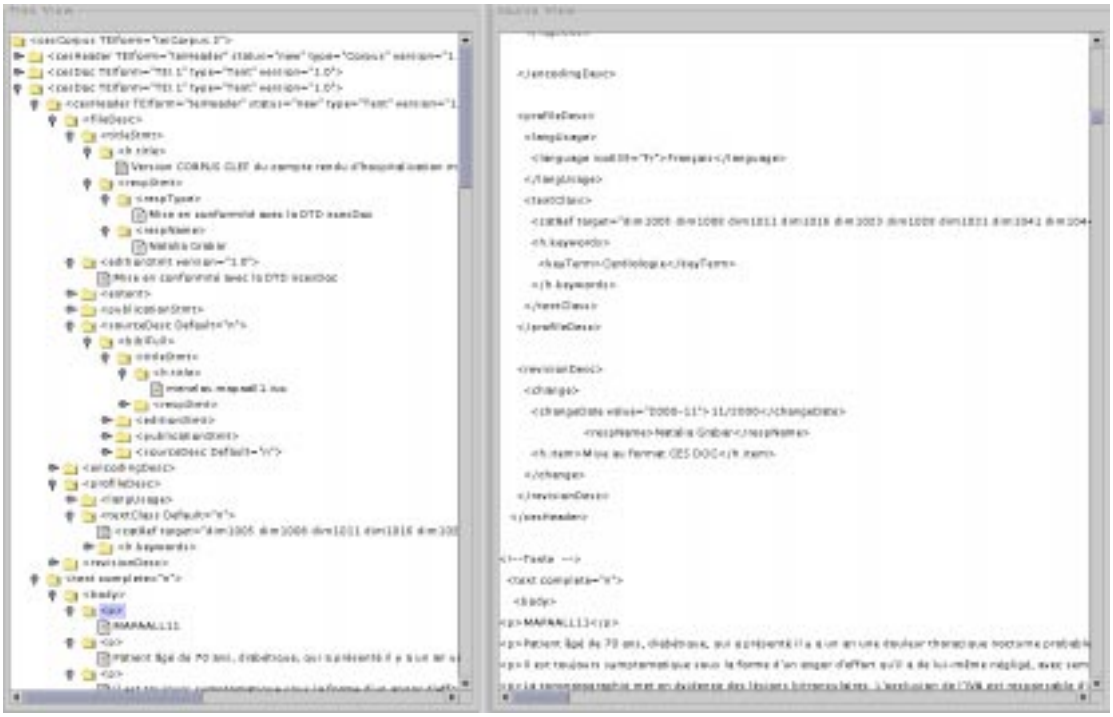

Figure 2: A slice of the implemented corpus: extracts of the document header of document 2 and first lines of its contents (viewed with Xerces TreeViewer).

As a proof of concept, we integrated 374 documents in the corpus: 294 anonymized patient discharge summaries from 4 different sites and 2 different medical specialties (cardiology, from project Menelas (Zweigenbaum & Consortium MENELAS, 1994), and haematology), 78 anonymized discharge letters, one chapter of a handbook on coronary angiography and one consensus conference on post-operative pain. The total adds to 143 Kwords, with an average of 385 words per document. Many colleagues have kindly declared their intent to contribute documents, so that a few million words should be attainable.

Adding new documents to the corpus and documenting them requires a varying amount of work depending on the type of document. Patient documents require the most attention because of anonymization. Their actual documentation also raises an issue: a precise documentation would re-introduce information on locations and dates, so that we must here sacrifice documentation for privacy.

A pre-specified model for document description is a need if a corpus is to be used by many different people. The dimensions of our model, implemented as taxonomic “categories”, will probably need some update with the introduction of the other main types of documents. We expect however that they should quickly stabilize.

4.3 Extracting sub-corpora

Adherence to an existing standard enabled us to implement our corpus model in a principled way with a very reasonable effort. Besides, the general move towards XML observed in recent years facilitates the conversion of existing documents and the subsequent manipulation of the corpus, which can be

manipulated through standard XML tools. We ran the Xerces Java XML library of the Apache XML project and James Clark's XT library under Linux, Solaris and HP-UX. The corpus was checked for syntactic well-formedness ("conformance") and adherence to the xcesDoc DTD ("validity").

We use XSL stylesheets to produce tailored summaries of the corpus contents and to extract sub-corpora. An XSL stylesheet can specify transformations that should be applied to an input XML file, here the whole corpus. These transformations include the selection of elements of the input file (here, individual texts) and the construction of a new document (a sub-corpus) embedding these elements. Additional material such as a new corpus header can be built on the fly as needed. Selection can operate on any of the features of the texts, including their documentation, so that all the previously discussed genres, dimensions and domains can serve as criteria for extracting corpus texts. We have written a few stylesheet to performs specific extractions. We are currently working on a generic user interface for specifying these extractions. An important need is to keep track of the origin of the corpus elements through successive extractions (Illouz *et al.*, 2000).

5 Conclusion and Perspectives

We have proposed a framework for designing a medical text corpus and a proof of concept implementation: a set of descriptive dimensions, a standardized encoding of both meta-information (implementing these dimensions) and content, and a small-size corpus of text samples encoded according to these principles.

This corpus, once sufficiently extended, will be useful for testing and training NLP tools: taggers, checkers, term extractors, parsers, encoders, information retrieval engines, information extraction suites, etc. We plan to distribute it to NLP and Medical Informatics researchers. We believe that the availability of such a resource may fill a gap in the current corpora and help better study the issues of genres in a specialized domain. The corpus should also allow more methodological, differential studies on the medical lexicon, terminology, grammar, etc.: *e.g.*, terminological variation across genres within the same medical specialty, or the correlation of observed variation with documented dimensions, which should teach us more about the features of medical language.

6 Acknowledgments

We wish to thank the French Ministry for Higher Education, Research and Technology for supporting project CLEF, D Bourigault and P Paroubek of project CLEF's management board for useful discussions, B Séroussi and J Bouaud for help about the document genres, and the many colleagues who agreed to contribute documents to the corpus.

Bibliography

Adda G, Mariani J, Paroubek P, Rajman M, Lecomte J 1999 Métrique et premiers résultats de l'évaluation GRACE des étiqueteurs morpho-syntaxiques pour le français. In Amsili P (ed), *Proceedings of TALN 1999 (Traitement automatique des langues naturelles)*, Cargèse, ATALA, pp 15–24.

Biber D 1994 Representativeness in corpus design. *Linguistica Computazionale*, IX-X:377–408. Current Issues in Computational Linguistics: in honor of Don Walker.

Biber D, Finegan E 1994 Intra-textual variation within medical research articles. In Ooostdijk N, de Haan P (eds), *Corpus-based research into language*, number 12 in Language and computers : studies in practical linguistics. Amsterdam, Rodopi, pp 201–222.

Darmoni S. J, Thirion B, Leroy J. P, Douyère M, Baudic F, Piot J 2000 CISMeF: a structured health resource guide for healthcare professionals and patients. In *Proceedings of RIAO 2000: Content-Based Multimedia Information Access*, Paris, France, C.I.D.

Dewe J, Karlgren J, Bretan I 1998 Assembling a balanced corpus from the internet. In *Proceedings of the 11th Nordic Conference on Computational Linguistics*, Copenhagen, pp 100–107.

Dolin R, Alschuler L, Boyer S, Beebe C 2000 An update on HL7's XML-based document representation standards. *Journal of the American Medical Informatics Association*, 7(suppl):190–194.

Dublin Core Metadata Inititative 1999 *The Dublin Core Element Set Version 1.1*. WWW page http://purl.org/dc/documents/rec-dces-19990702.htm.

Dunlop D 1995 Practical considerations in the use of TEI headers in large corpora. *Computers and the Humanities*, 29:85–98.

Friedman C 1997 Towards a comprehensive medical natural language processing system: Methods and issues. *Journal of the American Medical Informatics Association*, 4(suppl):595–599.

Giordano R 1995 The TEI header and the documentation of electronic texts. *Computers and the Humanities*, 29:75–85.

Grefenstette G 1994 *Explorations in Automatic Thesaurus Discovery*. Natural Language Processing and Machine Translation. London, Kluwer Academic Publishers.

Grishman R, Kittredge R (eds) 1986 *Analyzing Language in Restricted Domains*. Hillsdale, New Jersey, Lawrence Erlbaum Associates.

Habert B 2000 Des corpus représentatifs : de quoi, pour quoi, comment ? In Bilger M (ed), *Linguistique sur corpus : Études et réflexions*, volume 31 of *Cahiers de l'Université de Perpignan*. Presses universitaires de Perpignan, pp 11–58.

Ide N, Priest-Dorman G, Véronis J 1996 *Corpus Encoding Standard*. Document CES 1, MULTEXT/EAGLES, http://www.lpl.univ-aix.fr/projects/eagles/TR/.

Illouz G 1999 Méta-étiqueteur adaptatif : vers une utilisation pragmatique des ressources linguistiques. In Amsili P (ed), *Actes de TALN'99 (Traitement Automatique des Langues Naturelles)*, Cargèse, ATALA, pp 185–194.

Illouz G, Habert B, Folch H, Heiden S, Fleury S, Lafon P, Prévost S 2000 TyPTex: Generic features for text profiler. In *Proceedings of RIAO 2000: Content-Based Multimedia Information Access*, Paris, France, C.I.D., pp 1526–1540.

Lehmann D, de Margerie C, Pelfrêne A 1995 *Lecticiel – Rétrospective 1992–1995*. Technical report, CREDIF – ENS de Fontenay/Saint-Cloud, Saint-Cloud.

Rassinoux A.-M 1994 *Extraction et Représentation de la Connaissance tirée de Textes Médicaux*. Thèse de doctorat ès sciences, Université de Genève.

Rossi Mori A, Consorti F 1999 Structures of clinical information in patient records. *Journal of the American Medical Informatics Association*, 6(suppl):132–136.

Sager N, Friedman C, Lyman M. S (eds) 1987 *Medical Language Processing: Computer Management of Narrative Data*. Reading, Mass., Addison Wesley.

Scherrer J.-R, Lovis C, Borst F 1996 DIOGENE 2, a distributed information system with an emphasis on its medical information content. In van Bemmel J. H, McCray A. T (eds), *Yearbook of Medical Informatics '95 — The Computer-based Patient Record*. Stuttgart, Schattauer.

Sinclair J 1996 *Preliminary recommendations on Text Typology*. WWW page http://nicolet.ilc.pi.cnr.it/EAGLES/texttyp/texttyp.html, EAGLES (Expert Advisory Group on Language Engineering Standards).

Tuttle M, Olson N, Keck K, Cole W, Erlbaum M, Sherertz D, Chute C, Elkin P, Atkin G, Kaihoi B, Safran C, Rind D, Law V 1998 Metaphrase: an aid to the clinical conceptualization and formalization of patient problems in healthcare enterprises. *Methods of Information in Medicine*, 37(4-5):373–383.

Vihla M 1998 Medicor: A corpus of contemporary American medical texts. *ICAME Journal*, 22:73–80.

Wierzbicka A 1985 A semantic metalanguage for a crosscultural comparison of speech acts and speech genres. *Language in society*, 14:491–514.

Zweigenbaum P, Consortium MENELAS 1994 MENELAS: an access system for medical records using natural language. *Computer Methods and Programs in Biomedicine*, 45:117–120.