MEDINFO 2001
V. Patel et al. (Eds)
Amsterdam: IOS Press

Building a Text Corpus for Representing the Variety of Medical Language

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

[a] DIAM — Service d'Informatique Médicale/DSI, Assistance Publique – Hôpitaux de Paris
& Département de Biomathématiques, Université Paris 6
[b] LIMSI-CNRS & Université Paris 10

Abstract

Medical language processing has focused until recently on a few types of textual documents. However, a much larger variety of document types are used in different settings. It has been showed that Natural Language Processing (NLP) tools can exhibit very different behavior on different types of texts. Without better informed knowledge about the differential performance of NLP tools on a variety of medical text types, it will be difficult to control the extension of their application to different medical documents. We endeavored to provide a basis for such informed assessment: the construction of a large corpus of medical text samples. We propose a framework for designing such a corpus: a set of descriptive dimensions and a standardized encoding of both meta-information (implementing these dimensions) and content. We present a proof of concept demonstration by encoding an initial corpus of text samples according to these principles.

Keywords:

Natural language processing, text corpus, medical documents, French.

Introduction

Medical language processing has focused until recently on a few types of textual documents. Medical narratives, including discharge summaries and imaging reports, have been the most studied ones [1,2,3,4]. Short problem descriptions, such as signs, symptoms or diseases, have been the subject of much attention too, in relation to standardized vocabularies [5]. Some authors have also examined abstracts of scientific literature [6]. And indeed, web pages are today the most easily available source of medical documents. All these constitute different kinds of documents. They 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 [7]. The natural language processing (NLP) tools that have been tailored for one document type may therefore be difficult to apply to another type [2][1]. This 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 a basis for such informed assessment: the construction of a large corpus of medical text samples. We address this task for French, 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. 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 appropriate dimensions: origin, genre, domain, etc. These dimensions have to be documented precisely for each text sample. 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 oritinal formats of documents

We present here a framework for designing a medical text corpus: a set of 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 [10], 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 [2], whose goal is to build a large corpus of French text samples and to distribute it widely to researchers.

[1] The precision of French taggers evaluated within the framework of GRACE [8], measured in relation to a manually tagged reference corpus, similarly shows significant variations depending on the part of the corpus under examination [9].

[2] www.biomath.jussieu.fr/CLEF/

Background

Nowadays, for "general" language, "mega-corpora" are available, such as the BNC (*British National Corpus*) [11]: 100 million words (about 1,000 medium-size novels), comprising 10 million words of transcribed spoken English as well as written language. This corpus provides a set of textual data whose production and reception conditions are precisely defined and which is representative of a great variety of communication situations.

The available medical corpora we are aware of are collections of abstracts of scientific literature, *e.g.*, MEDIC, cited in [6]. Medical textbooks and scientific literature have been collected in project LECTICIEL [12] for French for Special Purposes learning. Users could add new texts to the database and compare them with the existing sub-corpora. One medical corpus was specifically built for the purpose of linguistic study: MEDICOR [13]. 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. 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. Very large collections of medical texts indeed exist within hospital information systems, the DIOGENE system being among the earliest ones [14]. The issue here is that of privacy and therefore anonymization, to which we return below.

Beyond bibliographic description, descriptive dimensions for characterizing text corpora have been proposed by Sinclair [15] and Biber [16] among others. A related strand of work is that around the standardization of meta-information for documenting web pages [17]; but this covers more limited information than that we shall need. In the medical informatics domain, the standardization efforts of bodies such as HL7 [18] and CEN [19] focus on clinical documents for information interchange: both their aim and coverage are different from ours.

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) and a Corpus Encoding Standard (CES) [10]. The CES DTD is therefore the natural format for encoding a corpus that is targeted at NLP tools.

Material and Methods

We explain in turn each of the main phases of the design of our corpus: (*i*) assessing document diversity and choosing dimensions to describe this diversity, *i.e.*, a kind of multi-axial terminology for describing textual documents, and (*ii*) implementing them in a standard XML DTD; then (*iii*) selecting the main classes of documents we want to represent and documenting them with these dimensions. We then explain how to populate the corpus with texts, and illustrate the method on currently integrated documents.

Studying and Representing Diversity

A large palette of medical textual documents are in use in different contexts. Our aim here is 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. 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. [20] 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 been shown [21] that, while there is no well-established genre palette for Internet materials, it is 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. So the very first step consists in asking people from the domain the main communicative routines or speech act they identify. We started 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. 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 communicative language. In order to gather a corpus, one must explicitly choose the language use(s) (s)he wants to focus on. The resulting variety is 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. Internal variety: a communicative routine is often associated with consistent stylistic choices, that is, observable restrictions in the choice of linguistic items: lexical items, syntactic constructions, textual organization, such as the standard four-part organization of experimental studies: Introduction, Methods, Results, Discussion[3]. Besides a given cluster of linguistic features can be shared between different communicative routines (for instance between discharge summaries and imaging reports).

We listed this way 57 different genres of medical texts. They include various reports (*e.g.*, discharge, radiology), letters (*e.g.*, discharge, referral), teaching material (*e.g.*, lecture notes), publications (*e.g.*, journals, books, articles), reference material (*e.g.*, encyclopedia, classifications, directories), guidelines (*e.g.*, recommendations, protocols), and official documents (*e.g.*, French Bulletin Officiel, code of deontology). These document types are difficult to

[3] Each of these parts was shown to have distinct linguistic features [7].

classify into non-overlapping groups. Therefore modelling the corpus with descriptive dimensions is all the more useful. To produce this set of dimensions, we first studied how the dimensions proposed in the literature covered differences in text types, and added to them as needed.

Within the TEI standardization group, much attention has been devoted to the definition of *headers* [22]. A header is a normalized way of documenting electronic texts. It describes the electronic text and its source (bibliographic information, when available), it gives the encoding choices for the text (editorial rationales, sampling policy...), non-bibliographical information that characterize the text, and a history of updates and changes. In the non-bibliographical part of the header, the text is described 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). It is then possible to extract sub-corpora following arbitrary complex constraints stated in these classification schemes. For instance, the interface to the BNC relies on such an approach [23] 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).

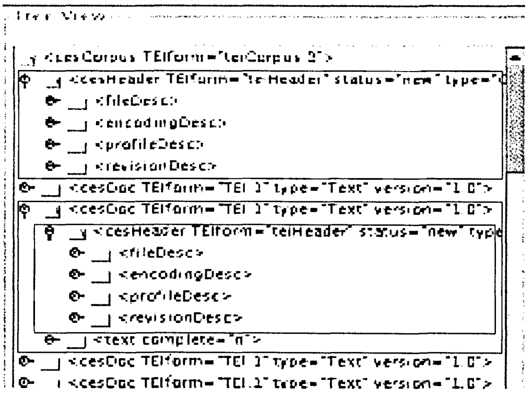
Implementing a Corpus Header

We checked whether our corpus model, with all its dimensions, could fit in the standard TEI XML CES model [10]. In the CES model, a corpus consists of a corpus header followed by a collection of documents, each of which is a pair of document header and text (figure 1). The corpus header caters for documenting the corpus as a whole, whereas each document header contains meta-information for its text. We could find a mapping into the CES header for each dimention of our model, and therefore implemented it in the CES framework. An added advantage is that the CES model provides additional documentation dimensions, *e.g.*, information about the corpus construction process (text conversion, normalization, annotation, etc.).

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 overall strategy of project CLEF is to opt for samples in the order of 2,000 words each. The expected benefits are a more manageable size and less trouble with property rights: it may be more acceptable for a publisher to give away extracts rather than full books or journals, so that text samples should be easier to obtain. The drawback is that textual phenomena with a larger span may not be studied on such samples. We thus plan to be flexible on sample size.

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). 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.



```
Tree View
<cesCorpus TEIform="teiCorpus.2">
  <cesHeader TEIform="teiHeader" status="new" type="
    <fileDesc>
    <encodingDesc>
    <profileDesc>
    <revisionDesc>
  <cesDoc TEIform="TEI.2" type="Text" version="1.0">
  <cesDoc TEIform="TEI.2" type="Text" version="1.0">
    <cesHeader TEIform="teiHeader" status="new" type
      <fileDesc>
      <encodingDesc>
      <profileDesc>
      <revisionDesc>
    <text complete="n">
  <cesDoc TEIform="TEI.2" type="Text" version="1.0">
  <cesDoc TEIform="TEI.2" type="Text" version="1.0">
```

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

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 prototypical information for texts of this type. This factorizes documentation work, so that the remaining work needed to derive suitable document headers for individual texts is kept to a minimum. Document templates were implemented for the text types included so far in the corpus.

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.

The document header template for the appropriate document type is then instanciated. For series of similar samples (*e.g.*, a series of discharge summaries), most of this instantiation can be performed automatically.

Results

The main results in the current state of the project are (*i*) a model of document description (the dimensions), (*ii*) an implementation of this model and (*iii*) the inclusion of a series of documents in this implementation (the current corpus).

We settled on 30 dimensions, partly derived from [15], [7] and [17]. The two main groups of dimensions are "external": bibliographic reference (*e.g.*, title, author, date; size and localisation of sample) and context of production (*e.g.*, institutional *vs* private, published or not, mode of production, of transmission, frequency of publication, source, destination). The dimensions of the last group are "internal": level of language, distance from readership, personalization of the message, factuality, technicity, style. Allowed values are specified for each dimension. One of the dimensions is the domain of the text, here the medical specialties involved. We reused and slightly adapted the list of domains that help to index medical web sites on the CISMEF directory (www.cismef.org).

The implemented model fits as an instance of the XML CES DTD (xcesDoc.dtd) (www.cs.vassar.edu/XCES/). Bibliographical dimensions are explicitly modelled in that DTD within each document header. For dimensions pertaining to the context of production and internal dimensions, "taxonomies" are defined in the corpus header: they consist of hierarchies of category descriptions. Each document in the corpus is characterized by a set of such categories: this is implemented by referring to these standard categories in the "profile description" section of that document's header. Figure 2 shows a slice of the implemented corpus.

As a proof of concept, we integrated 374 documents in the corpus: 294 patient discharge summaries from 4 different sites and 2 different medical specialties (cardiology, from project Menelas [4], and haematology), 78 discharge letters, one chapter of a handbook on coronary angiography and one "conference of consensus" 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.

The corpus 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.

Discussion

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. A few lines of XSL instructions suffice to design extraction methods which are then executed in seconds on the whole corpus.

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.

The XCES DTD was designed to cope with multilingualism, including for non-western languages and scripts. It caters for language declarations at every level of granularity. This facilitates the extension of the corpus to multiple languages or the parallel development of corpora for different languages based on a common model.

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, robust parsers, encoders, information retrieval engines, information extraction suites, etc. We plan to distribute it to Medical Informatics and NLP researchers. We believe that the availability of such a resource may be an incentive to attract more generalist NLP researchers to work on medical texts. The corpus will 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.

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.

References

[1] Sager N, Friedman C, and Lyman MS, eds. *Medical Information Processing - Computer Management of Narrative Data.* Addison Wesley, Reading Mass, 1987.

[2] Friedman C. Towards a comprehensive medical natural language processing system: Methods and issues. *J Am Med Inform Assoc* 1997;4(suppl):595–9.

[3] Rassinoux AM. *Extraction et Représentation de la Connaissance tirée de Textes Médicaux.* Thèse de doctorat ès sciences, Université de Genève, 1994.

[4] Zweigenbaum P and Consortium MENELAS . MENELAS: an access system for medical records using natural language. *Comput Methods Programs Biomed* 1994;45:117–20.

[5] Tuttle M, Olson N, Keck K, et al. Metaphrase: an aid to the clinical conceptualization and formalization of patient problems in healthcare enterprises. *Methods Inf Med* November 1998;37(4-5):373–83.

[6] Grefenstette G. *Explorations in Automatic Thesaurus Discovery.* Kluwer, London, 1994.

[7] Biber D and Finegan E. Intra-textual variation within medical research articles. In: Ooostdijk N and de Haan P, eds, *Corpus-based research into language*, number 12. Rodopi, Amsterdam, 1994:201–22.

[8] Adda G, Mariani J, Paroubek P, Rajman M, and Lecomte J. Métrique et premiers résultats de l'évaluation GRACE des étiqueteurs morpho-syntaxiques pour le français. In: Amsili P, ed, Actes de TALN 1999, Cargèse. July 1999:15–24.

[9] Illouz G. Méta-étiqueteur adaptatif : vers une utilisation pragmatique des ressources linguistiques. In: Amsili P, ed, Actes de TALN 1999, Cargèse. July 1999:185–94.

[10]Ide N, Priest-Dorman G, and Véronis J. Corpus encoding standard. Document CES 1, MULTEXT/EAGLES, http://www.lpl.univ-aix.fr/projects/eagles/TR/, 1996.

[11]The British National Corpus. http://info.ox.ac.uk/bnc/, Oxford University Computing Services, 1995.

[12]Lehmann D, de Margerie C, and Pelfrêne A. Lecticiel – rétrospective 1992–1995. Technical report, CREDIF – ENS de Fontenay/Saint-Cloud, Saint-Cloud, 1995.

[13]Vihla M. Medicor: A corpus of contemporary American medical texts. *ICAME Journal* 1998:73–80.

[14]Scherrer JR, Lovis C, and Borst F. DIOGENE 2, a distributed information system with an emphasis on its medical information content. In: van Bemmel JH and McCray AT, eds, *Yearbook of Medical Informatics 95.* Schattauer, Stuttgart, 1996.

[15]Sinclair J. Preliminary recommendations on text typology. Technical report, EAGLES (Expert Advisory Group on Language Engineering Standards), june 1996.

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

[17]The Dublin Core element set version 1.1. WWW page http://purl.org/dc/documents/, Dublin Core Metadata Inititative, 1999.

[18]Dolin R, Alschuler L, Boyer S, and Beebe C. An update on HL7's XML-based document representation standards. In: Proc AMIA Symp, 2000:190–4.

[19]Rossi Mori A and Consorti F. Structures of clinical information in patient records. In: Proc AMIA Symp, 1999:132–6.

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

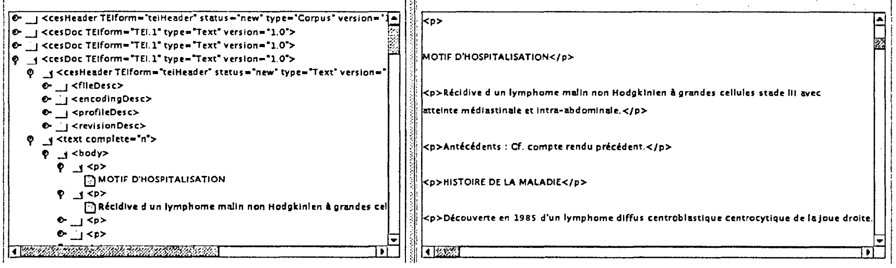
[21]Dewe J, Karlgren J, and Bretan I. Assembling a balanced corpus from the internet. In: 11th Nordic Conference on Computational Linguistics, Copenhagen. 1998:100–7.

[22]Giordano R. The TEI header and the documentation of electronic texts. *Comput Humanities* 1995(29):75–85.

[23]Dunlop D. Practical considerations in the use of TEI headers in large corpora. *Comput Humanities* 1995(29):85–98.

Address for correspondence

Pierre Zweigenbaum DIAM — SIM/DSI/AP-HP
91, boulevard de l'Hôpital, 75634 Paris Cedex 13, France
pz@biomath.jussieu.fr, http://www.biomath.jussieu.fr/ pz/



```
<cesHeader TEIform="teiHeader" status="new" type="Corpus" version="
<cesDoc TEIform="TEI.1" type="Text" version="1.0">
<cesDoc TEIform="TEI.1" type="Text" version="1.0">
<cesDoc TEIform="TEI.1" type="Text" version="1.0">
  <cesHeader TEIform="teiHeader" status="new" type="Text" version="
    <fileDesc>
    <encodingDesc>
    <profileDesc>
    <revisionDesc>
  <text complete="n">
    <body>
      <p>
        MOTIF D'HOSPITALISATION
      <p>
        Récidive d un lymphome malin non Hodgkinien à grandes cel
      <p>
      <p>
```

```
<p>

MOTIF D'HOSPITALISATION</p>

<p>Récidive d un lymphome malin non Hodgkinien à grandes cellules stade III avec
atteinte médiastinale et intra-abdominale.</p>

<p>Antécédents : Cf. compte rendu précédent.</p>

<p>HISTOIRE DE LA MALADIE</p>

<p>Découverte en 1985 d'un lymphome diffus centroblastique centrocytique de la joue droite.
```

Figure 2: A slice of the implemented corpus: the first lines of document 4 (viewed with Xerces TreeViewer).