Ceusters W. Language, Medical Terminologies and Structured Electronic Patient Records: how to escape the Bermuda Triangle. In De Moor G, De Clercq E (eds.) Proc 18th MIC Conf, 2000;:7-14.

Language, medical terminologies and structured electronic patient records: how to escape the Bermuda triangle.

Dr. W. Ceusters

Abstract. Structured electronic patient record systems generally embed controlled vocabularies and terminologies as a (naïve) mechanism to make users structure manually in a machine readable format *clinical data* that they otherwise would register as *clinical narrative*. In this paper, we argue that modern language engineering applications have become mature enough to make the traditional approach gradually obsolete. Medical terminologies, including coding and classification systems will still fulfil specific roles, though different ones than before. Structuring narrative data will not be a user responsibility, but a system responsibility.

1. Introduction

Though most clinicians and other healthcare workers are gradually becoming convinced of the advantages of using computers, they still prefer to retrieve data stored by others, than to register data themselves. There are many reasons for this such as unavailability of systems at the point of care, incomplete integration in the primary care process, or the fact that only a subset of the activities for which clinicians would like to have computer support, are actually offered.

The issue that deserves our particular attention in this paper is the *information structuring bottleneck*. Healthcare records, whether on paper or in computers, are originally kept as an external record for individual patient histories, such that future decisions can be based appropriately on past events. Electronic patient record systems have additional advantages over paper-based systems in their ability to allow for cross-patient studies, and to provide active decision management functionalities. While the former requires thorough structuring of the data inside the machine, the latter also requires representing and storing knowledge and information in the machine so that the machine *itself* can manipulate it, at least for tasks for which it is better suited than humans.

The need for structured *data representation and storage* being undeniable and very well understood, the need for structured *data entry* seems to be the logical consequence. This is at least the impression that we get from analysing the data acquisition interfaces of so many electronic healthcare record systems. There is structuring at the level of the data capture modalities such as rigorous data entry forms, point and click interfaces, structured menu's, etc. There is also structuring at the level of content by using coding and classification systems or controlled vocabularies. The question should be whether or not it is necessary to require the structuring be done by the user. Or as Tange et al. phrase it: "*Initiatives to facilitate the entry of narrative data have focused on the control rather than the ease of data entry*" ([1], p. 24). It is a fact, that most users don't like structured data entry at all, but that many accept it in the light of the benefits obtained when retrieving

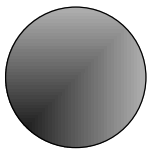
information. They accept the burden of structured data entry as the price to be paid for powerful information retrieval. But is this price affordable, let alone justifiable ? Many clinicians share the view that faithful recording of patient data can only be achieved by using natural language. This was already stated in the early eighties by Wiederhold who claimed that *the description of biological variability requires the flexibility of natural language and it is generally desirable not to interfere with the traditional manner of medical recording* [2]. Also more recently, strong arguments have been given to preserve natural language registrations in clinical records and to view them under a "narratological framework" as proposed by Kay and Purves [3].

Besides this theoretical and fundamental position in favour of natural language registration, there is also a practical reason: data entry by means of continuous speech recognition (CSR). CSR technology has now reached a functional threshold in transforming a speech signal into digital text what is all that is needed for dictation. However, inexperienced users quickly might infer from this evolution that all data entry could be done by voice, freeing them from the need to use a keyboard. Despite this demand, CSR is not that easy lined up with structured data entry forms or cascaded menu's. The command and control paradigm for navigating through forms and menu's is only acceptable in a "hands free" situation, but even that still requires visual feedback from the screen. The ideal situation would be one in which users can enter information or issue queries in natural language, upon which the machine would analyse and structure the input automatically. This calls for advanced natural language understanding.

2. A geometrical view on clinical data entry

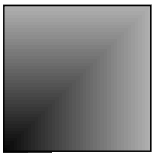
The three components we deal with in this paper are clinical language, medical terminologies (ranging from simple coding and classification systems to formal ontologies) and electronic patient records which respectively can be symbolised by a circle, a triangle and a square.

Clinical language is symbolised by a circle for its ability of smooth representation of ideas relatively independent of a rigid formalism: the same ideas can be expressed in various ways, with plenty of room to specify small but relevant details. Language is also very dynamic: it tends to change over time or specific phrasings are often influenced by local conditions, in the same way that a circle (or better a sphere) can turn gently around when touched. However, when pushed too hard, the sphere can move too far away from its original position, just as improper language use by the clinical narrator may result in ambiguous interpretation afterwards.

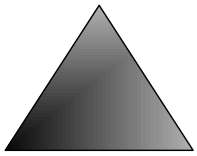


Clinical Language

Electronic healthcare record systems (EHCRs) at the other hand are designed for stability, and hence are symbolised by a square. The square symbolises also the rigid internal structure of most EHCRs. This is necessary for making data useful for subsequent automated processing, but leads too often to "squarely" structured user-interfaces that don't meet users' demands for completeness and adequate subtlety of expression.



EHCRS



Terminologies

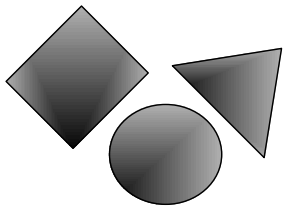
Finally terminologies can best be thought of as triangles representing a focused thrust forward from a strong base. Medical terminologies fulfil basically two goals. They try to fix domain semantics such that the meaning of terms changes little over time: they (can) provide a solid and stable base for unambiguous data registration. Most systems, especially coding and classification systems, are designed with a specific purpose in mind such as mortality and morbidity statistics, or reimbursement: hence they are focused.

3. Widely used paradigms for clinical data entry

Implementations of electronic patient record systems should find an adequate balance in dealing with clinical language, rigid database structures and medical terminologies: they should be able to fit the circle, the square and the triangle into a harmonic and harmonious whole. Unfortunately, this is not the case with the clinical data entry paradigms most systems adhere to today.

Many papers describe the kind of data that are to be registered in an EHCR, some from a standardisation perspective [4], others on more technical or scientific grounds [5]. Prior to define a framework for modelling the EHCR, a clinical account is given by Rector et al [6, 7]. An essential criteria is that the record should give a faithful account of the clinician's understanding. Data should be formulated in terms that are found natural. Conflicting statements must be allowed and also uncertain and negative statements must be accepted. Descriptions should be given at any arbitrary level of detail and at the clinicians' natural level of abstraction. Once entered, data should be there permanent. Though this description fits the characteristics of free text registration, the authors argue that also structured data entry paradigms should fulfil these requirements. Unfortunately, they never do.

3.1 Systems suffering from the "ICPC syndrome"



"ICPC syndrome"

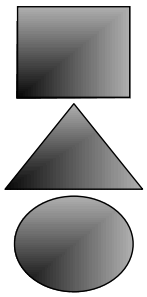
The ICPC (International Classification of Primary Care, currently being replaced by ICPC-2) has proved to be a valuable tool for statistically comparing the activities of GP surgeries, based around the concept of "reason for encounter". It consists of a small classification of around 780 terms that clearly cannot be used to describe all relevant information with respect to individual patient care. The same can be said of other coding and classification systems that try to generalise healthcare information by abstracting away from the details that are judged irrelevant for the specific purpose that they have been designed for. But irrelevant for a specific purpose, does not necessarily mean irrelevant for all individual patients.

The logical conclusion is: if systems are not designed for capturing all relevant data for individual patient care, then don't use them for that purpose ! Hence, developers of electronic patient record systems that want to integrate these systems (usually only available as long lists without adequate searching facilities) into their applications have only one good option: the systems must be integrated in addition to other data entry facilities, and users must be instructed that it does not suffice to register a number of codes out of such systems to have a faithful recording. They'll

have to register in free text, and then must assign codes afterwards for all systems that are required according to institutional or governmental directives. Only medical natural language understanding technology can improve this situation [8 - 11].

3.2 Use of controlled vocabularies

Controlled vocabularies are (possibly hierarchically) structured sets of certified terms that are verbal canonical representations of concepts. The aspect of control in a controlled vocabulary is related to the position of a specific term in the vocabulary as a whole, the choice of a particular term as canonical form, and the requirement that only terms from within the vocabulary are to be used in an application. They most often are implemented as "picking lists" at the level of the user interface.



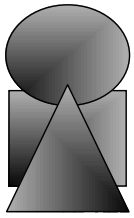
Controlled vocabulary

Controlled vocabularies are difficult to use. The lists tend to be too short such that adequate details cannot be recorded, or too long, such that it becomes impossible to find what one looks for. When properly integrated in EHCRs, they can account for some form of balance between the three components, though the balance is difficult to maintain.

3.3 Predictive data entry

Traditional terminologies (nomenclatures, thesauri, classifications, etc.) are designed to be used by humans. Even electronic versions of these systems, in which it is possible to browse through the hierarchies of the terminology, are still intended to be used by humans, the computer just being there as a replacement for the book. A major problem for such naïve electronic versions is that they cannot take advantage of the knowledge implicitly available in the terms (or the rubrics in classification systems), but that they must rely on the limited knowledge available in the generic links between terms. Finding specific terms requires a priori knowledge by the user on how the system is structured. With flat terminologies, in which large quantities of narrower-terms depend from one broader-term, the computer is even seen as a burden, because only a limited number of terms can be seen at the same time on the screen. A second disadvantage is that the terminologies only can be viewed in their original structure, and that reclassification of the terms, following different criteria, cannot be realised.

In order to overcome these problems, terminologies must be expressed in a formal way: all the knowledge must be made explicit such that machines can exploit it autonomously. Major efforts have been conducted in this respect, PEN&PAD and GALEN being the most notable of it [12]. Such models have been shown to allow dynamic, real-time generation of controlled vocabularies that "pop-up" depending on the information that has been entered before. Interfaces driving upon this paradigm allow for "predictive data entry": the machine predicts what the user "sensibly" will enter next. Unfortunately, what can be entered, depends again on the exhaustiviness of the vocabulary (or better: the concept model underneath it).



Predictive data entry

4. Medico-linguistic ontologies

In our view, faithful registration of patient data can only be achieved when using natural language. Given the need for structured representations, we argue that structuring must not be done by the user, but by the machine. This can only be achieved by natural language understanding (NLU) applications that are to be integrated as middleware components in EHCRs.

NLU requires more than terminologies alone. Formal terminologies are a start, provided that they are built keeping NLU in mind. If not, they are doomed to fail. When formalising terminologies, an *ontology* has to be defined, i.e. a representation - to be used in computer systems - of what concepts exist in the world, and how they relate to one another. Ontologies are often viewed as strictly language independent models of the world, especially in the medical informatics community [13]. Unfortunately, models designed from this perspective cannot be used without great difficulties to understand medical language as we predicted in [14] and proved in [15].

This calls for formal terminologies that are built along three dimensions: a cognitive one, a linguistic one, and a communicative one. It has been shown that ontologies that are developed for solving particular problems in knowledge based applications are better suited to assist language understanding when the concepts and relationships (conceptual dimension) they are built upon, are also linguistically motivated [16]. Linguistic semantics based analyses allow us to separate f.i. entities from events and property concepts, a rather crude distinction being the fact that in most languages these concepts are respectively grammaticalised by means of nouns, verbs and adjectives [17]. Linguists are concerned on how these concepts give overt form to language, while from a computational point of view, these concepts also have to be "anchored" in a *linguistic ontology*.

While working on the language engineering aspects of Galen-In-Use, numerous examples were found where linguistic principles were in conflict with conceptual principles [18]. Physicians wants to see medical concepts organised in a framework that reflects their clinical way of thinking. As an example, the Galen model categorises the concepts of "filling" and "injecting" as specialisations of a "LiquidInstallingProcess" that itself is a child of "InstallingProcess". This categorisation is useful from a clinical perspective where from the place in the hierarchy it can be derived that the concepts of injecting and filling have to do with the installation of liquid (though not necessarily exclusively as the Galen model supports multiple parents). This categorisation does however not line up with the linguistic structures that (at least in European languages) are used to express installing, filling and injecting events. From a language understanding perspective, it would be better to categorise these motion events according to the way the thematic roles of *goal* and *theme* may surface in sentences expressing these events.

Finally, the communicative dimension of terminologies is both related with the maintenance of terminologies, and the purpose(s) for which they are designed. As a consequence, problems such as how to guarantee that a (formal) terminology is properly used for what it is designed for, how can it be put in practice, how can it be maintained, and what is needed to allow co-existence with other systems, need to be accounted for. To all these questions, there is one common answer: there must be a general computational framework upon which various terminological tools and applications can be built, and that has formal links to natural languages, medical

terminologies, and electronic healthcare record systems (see [19] for a detailed account in the domain of pharmaceutical medicine).

5. Towards harmony in clinical system design

These are the facts that (in our view) dictate EHCR system design from this day on:

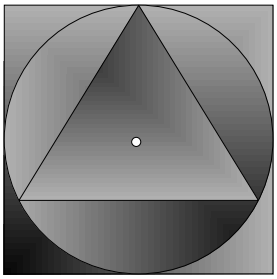
1) natural language is the only medium that is able to communicate clinical information about individual patients without loss of necessary detail;

2) structured data repositories are required to make subsequent analyses possible;

3) any transformation from free language to coding and classification systems results in information loss that is unacceptable for individual patient care, but at the other hand is a conditio sine qua non for population based studies;

4) today's graphical user interfaces can deal reasonably well with picking lists build around controlled vocabularies that fulfil a bridging function from free language towards coding and classification systems. However, speech recognition technology will soon free the user from the screen, such that item selection isn't anymore an option.

5) User interfaces must be designed in such a way that they don't disturb the primary process. There must come a shift from the current paradigm of user-initiated "data-entering" towards machine-initiated "data-capture": the machine observes without any interference of what is going on.



To make this happen, medico-linguistic ontologies will need to become essential components of any EHCR system. Medical ontologies that have been designed without keeping the language-constraints in mind, are doomed to fail: "*The current implementation of SNOMED-RT does not have the depth of semantics necessary to arrive at comparable data or to algorithmically map to classifications such as ICD-9-CM*" [20, p70], or also "*A serious limitation of the Galen approach is that specialisation is invariably linked to a conceptual relation*" [21, p66]. The same goes for systems that are mainly build around language, without adequate conceptual design, such as is the case for UMLS and its components: "*Simply using everything in the Metathesaurus does not make a good coding system*" [22], and "*The problems with the Metathesaurus as a single monolithic vocabulary are: 1. There is a wide range of granularity of terms in different vocabularies, 2. The Metathesaurus itself has no unifying hierarchy, so you cannot take advantage of hierarchical relations, 3. There may be other features of vocabularies that get lost in their "homogenisation" upon being entered into the Metathesaurus.*" [23].

The only good approach is to have systems that keep natural language, structured representations and formal terminologies nicely in balance. If clinicians understand well Francis Bacon's saying "*He who will not apply new remedies, must expect old evils*", why wouldn't EHCR systems developers do the same ?

6. References

[1] Tange HJ, Hasman A, de Vries Robbe PF, Schouten HC. Medical narratives in electronic medical records. International Journal of Medical Informatics, 1997, 46(1): 7-29.

[2] Wiederhold G. Databases in healthcare. Stanford University, Computer Science Department, Report No. STAN-CS-80-790, 1980.

[3] Kay S, Purves IN. Medical Records and Other Stories: a narratological framework. Methods of Information in Medicine 1996; 35: 72-87.

[4] Gabrielli ER. Standards for Electronic Patient Records. Journal of Clinical Computing, 20 (1), 1991, 21 - 32.

[5] van Ginneken AM, tam H, Moorman PW. A multi-strategy approach for medical records of specialists. International Journal of Biomedical Computing 42: 1996, 21-26.

[6] Rector AL, Nowlan WA, Kay S. Foundations for an electronic medical record. Meth Inform Med 30: 1991, 179-186.

[7] Rector AL, Nowlan WA, Kay S, Goble CA, Howkins TJ. A framework for modelling the Electronic Medical Record. Meth Inform Med 32: 1993, 109-119.

[8] Ceusters W, Laga M. *Introducing Language Engineering Tools to Support Information Processing in Healthcare Telematics*. In: Proceedings of Toward an Electronic Health Record Europe '99, 14-17 November 1999, London (UK), 251-255, 1999.

[9] Ceusters W. *Harmonisation and Formalisation of Nursing Terminology: a three-dimensional approach.* In RA Mortensen (ed.) ICNP and Telematic Applications for Nurses in Europe, IOS PRESS Amsterdam, 164-173, 1999.

[10] Ceusters W. *Talking to computers: natural language understanding for user-friendly interfaces*. Proceedings of the Third Convincing Cases in Healthcare Informatics Conference, Thessaloniki, Greece, 10-12/12/1999 (in press).

[11] Ceusters W, Lorré J, Harnie A, Van Den Bossche B. *Developing natural language understanding applications for healthcare: a case study on interpreting drug therapy information from discharge summaries*. Proceedings of IMIA-WG6, Medical Concept and Language Representation, Phoenix, 16-19/12/1999, 124-130.

[12] Rector AL, Glowinski A, Nowlan WA, Rossi-Mori A. Medical concept models and medical records: an approach based on GALEN and PEN&PAD. *Journal of the American Medical Informatics Association* 1995, 2: 19-35.

[13] Rector AL, Rogers JE, Pole P. The GALEN High Level Ontology. In Brender J, Christensen JP, Scherrer J-R, McNair P (eds.) *MIE 96 Proceedings*. Amsterdam: IOS Press 1996, 174-178.

[14] Ceusters W, Deville G, Buekens Ph. *The Chimera of Purpose- and Language Independent Concept Systems in Health Care*. In Barahona P, Veloso M, Bryant J (eds.) Proceedings of the XIIth International Congress of EFMI, 1994, 208-212.

[15] Ceusters W, Rogers J, Consorti F, Rossi-Mori A. *Syntactic-semantic tagging as a mediator between linguistic representations and formal models: an exercise in linking SNOMED to GALEN*. Artificial Intelligence in Medicine 1999; 15: 5-23.

[16] Mahesh K & Nirenburg S. A situated ontology for practical NLP. In *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95*. Montreal, Canada, 1995.

[17] Frawley W. Linguistic Semantics. Hilsdale, Hove and London: Lawrence Erlbaum Associates, 1992.

[18] Ceusters W. *Language Engineering as an Enabling Technology for Clinical Terminology Harmonisation*. In: CEC-DGXIII (ed.) Important Issues in Today's Telematics Research, TAP'98 Conference Barcelona, 1998, 168-173.

[19] Ceusters W, Bouquet L. Language Engineering and Information Mapping in Pharmaceutical Medicine. MIM-News vol 7 nr 1, 26-34, 2000.

[20] Elkin PL, Harris M, Ogren PV, Buntrock ID, Brown SH, Solbrig HR, Chute CG: "Semantic Augmentation of Description Logic based Terminologies" Addendum to Proceedings of IMIA-WG6, Medical Concept and Language Representation, Phoenix, 16-19/12/1999, 70-81.

[21] Hahn U, Schulz S, Romacker M: Part-whole reasoning: a case study in medical ontology engineering". IEEE Intelligent Systems & Their Applications vol 14 nr 5, 1999: 59-67.

[22] William T. Hole M.D., Director, Metathesaurus Research and Development, National Library of Medicine. Message to UMLS-users Mailing List, 23-06-2000.

Ceusters W. Language, Medical Terminologies and Structured Electronic Patient Records: how to escape the Bermuda Triangle. In De Moor G, De Clercq E (eds.) Proc 18th MIC Conf, 2000;:7-14.

[23] Hersh W. Message to UMLS-users Mailing List, 23-06-2000.