The myth of preferred terms in medical sublanguage and its impact on natural language understanding applications: an empirical study.

C. Moerkerke [1], W. Ceusters [2]

[1] *Mercator Hogeschool Gent, Dept Vertaalkunde*

[2] *Language & Computing nv, Zonnegem, www.landc.be*

Abstract. Clinical terminologies are usually build around "preferred terms" that are claimed to avoid misunderstandings when used for clinical registration and subsequent communication. This study gives indications that (at least in the domain of oto-rhino-laryngeology) misunderstandings are very unlikely to occur when non-preferred terms are used in discharge summaries and out-patient consultation reports. In addition, the study reveals that preferred terms (in the same domain) do not provide a good basis for medical natural language understanding applications, but that only corpus based approaches can account for the variability in language use by clinicians.

1. Introduction

Two reasons are often put forward to prefer coded medical data over free text: the ambiguity of natural language on the one hand, and the need for structured data to allow processing by machines on the other hand. While the latter is undeniably true (though in our view natural language is also "structured" and hence can also be processed by machines if the right algorithms are applied, e.g. [1]), the former is very much debatable, especially in well-defined domains such as healthcare. Examples often given are the existence of homonyms (terms with different meanings) such as "tuba dysfunction" in which case it is impossible to know whether dysfunction of the ovary duct or of the Eustachian tube is meant. But one seems too easily to forget that terms are very seldom used in isolation, or without any context. Also, "difficult to understand", is not the same as "impossible to understand". Cognitive scientists have conducted many experiments that prove that people require more time to analyse "ambiguous" sentences, but that this does not prevent them from coming to the right conclusions [2].

Anyway, as a solution to overcome problems of "ambiguous wordings", the medical informatics literature proposes controlled vocabularies that consist of "preferred terms", i.e. terms that are claimed to represent the meaning behind them in an unambiguous way, and that are accepted as such by the (clinical) community for which they are selected. Questions that inevitably rise are how many terms in real life really introduce understanding problems (is there a problem at all ?), and to what extend do clinicians accept terms as being preferred or not (can preferred terms solve the problem?).

The work presented in this paper is not the result of a clearly focused and well designed study intended to give answers to these questions. On the contrary: the data became available by coincidence due to some problems in translating a set of terms from Dutch to Spanish. The absence of a study methodology requires us to be cautious when interpreting some findings. Where that is the case, relevant methodology-related questions (MRQ) are given as end-notes to this paper. But nevertheless, the results obtained put claims related to the usefulness of preferred terms in a different perspective.

2. Material and methods

A corpus of 7.605 discharge summaries and out-patient consultation reports coming from 6 different oto-rhino-laryngeologists was processed by an automated language-independent statistical term extractor to extract meaningful terms [3]. The terms were ranked according to their frequency in the corpus and annotated as belonging to categories such as symptoms, diseases, procedures, etc. A student in translation studies was given the task to verify whether or not the 60 most frequent terms could be found as "preferred" terms in primary and secondary sources, and then to provide the most adequate Spanish translation for these terms. This was realised as part of her thesis [4].

Early during the work it became apparent that many terms could not literally be found in authoritative sources in the domain. As a consequence, it was decided to perform the translation work on similar terms (with the same meaning as the original ones) on the basis of the literature that was available.

An interesting question that came up was whether or not the terms not literally found in authoritative sources (but remember: with high frequency occurring in the corpus !) had to be considered as depreciated terms, or terms that only locally were used by the clinicians whose reports were collated in the corpus. For this reason, a questionnaire was sent to 235 ORL clinicians, with for each term the following questions (amongst a few other that are not relevant for this paper):

1) Do you know this term ? (y/n)
2) Do you use this term never/sometimes/often/very often/always ?
3) What term would you rather use in your reports ?

The terms asked to comment upon were a mixture of those found in the original reports, and those suggested as similar terms found in the literature.

3. Results

110 clinicians returned the questionnaire before the deadline.

Table 1 shows the results for the terms found in the original reports, but not in the literature. The following figures are displayed:

- Column (1): Frequency of the term in the corpus. Note that this frequency is influenced by both the incidence of the symptoms in clinical practice, and the particular preferences of the clinicians that contributed to the corpus to use these terms instead of other similar ones.
- Column (2): Percentage of the respondents indicating that they know the meaning of the term.
- Columns (3) - (7): Number of respondents that indicated to use the term never, rarely, often, very often, or always respectively.
- Column (8): calculated measure reflecting the overall indicated use of the term by all respondents. The following formula was used:

100 * (C4 + (2 * C5) + (3 * C6) + (4 * C7)) / M * (C3+C4+C5+C6+C7)

where Cx denote the respective column, and M was set to 2.8 such that this measure returned 50 for the overall use of all terms by all respondents. As a consequence, results lower than 50 for individual terms indicate less frequent use as compared to the overall use.

	N	Fam			Use			
			never	rarely	often	very often	always	M
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
aspecifieke mucosahyperreactiviteit	55	96	19	36	18	5	3	44
klassieke inhalatie-allergenen	94	96	8	32	29	4	9	60
acute exacerbaties	203	100	29	39	7	7	1	34
livide slijmvliezen	44	65	54	17	0	0	1	10
congestieve neusmucosa	31	99	7	15	29	25	7	76
hypertrofe onderste neusschelpen	113	100	9	8	38	24	3	73
habituele neusademhaling	37	88	28	17	20	6	5	45
habituele mondademhaling	63	98	18	22	28	10	3	53
nasale klank	143	99	9	34	19	17	3	59
moeilijke neusademhaling	94	95	20	22	27	9	1	48
etmoidale sluiering	67	100	5	13	42	16	6	74
etterige secreties	46	99	3	14	26	29	5	80
obstructief ademen	33	88	25	19	28	5	2	44
crypteuse amandelen	75	99	8	16	25	25	6	74
kissing tonsils	43	99	14	40	7	15	1	48
dysphagie voor vloeistoffen	36	99	7	35	22	11	4	58
dysfonie bij intensieve stembelasting	92	98	15	38	15	11	1	47
onvolledige stembandsluiting	18	100	6	23	30	13	4	65
epitympanale trommelvliesretracties	21	96	30	26	12	9	1	37
auditieve communicatievaardigheid	81	85	47	27	1	0	1	16
discreet transmissieverlies	57	99	25	23	15	13	1	45
neurosensorieel verlies	212	96	7	27	24	13	6	64
echte draainissen	70	89	43	16	12	5	0	26
carotis souffle	30	99	14	49	9	6	1	40
faciale pijnen	54	98	11	39	18	8	1	48
maxillaire tandpijn	66	84	33	30	10	3	1	29
opgezette halsbasis	24	80	37	26	8	3	1	26
koude nodulus	23	96	5	23	24	15	10	72
TOTALS			536	726	543	307	88	50

Table 1: Use of some medical terms by oto-rhino-laryngeologists. See text for details.

Table 2 shows the same type of results for the terms that were proposed by the students as a replacement for the terms out of table 1 that could not be found in authoritative sources. Obviously, column 1 is left blank as these terms were not found in the original corpus.

	N	Fam			Use			
	(1)	(2)	never (3)	rarely (4)	often (5)	very often (6)	always (7)	M (8)
niet-specifieke hyperreactiviteit		99	19	30	19	8	1	45
inhalatie-allergenen		100	2	27	26	22	4	71
exacerbatie		99	19	35	13	10	1	43
gezwollen neusslijmvlies		100	6	20	32	19	2	67
neusademhaling		99	1	16	23	25	15	88
habitueel mondademen		99	3	27	30	12	5	66
nasale spraak		100	5	20	22	27	6	75
gestoorde neusademhaling		96	13	26	27	14	0	54
sluiering van de sinus etmoidalis		100	4	23	32	19	1	67
purulente afscheiding		100	8	31	23	13	4	60
belemmerde ademhaling		95	11	30	18	18	0	56
crypteuse tonsillen		99	12	16	24	18	10	71
echte draaiingen		99	41	27	4	7	0	25
stemmisbruik		100	1	13	17	40	9	91
onvolledige sluiting vd stembanden		100	4	16	26	31	2	76
perceptieverlies		100	2	11	20	35	9	89
carotis geruis		99	15	47	7	10	0	41
aangezichtsspijnen		100	5	31	18	20	6	67
maxillaire pijn		100	11	28	25	12	3	57
koude nodus		99	26	24	21	8	4	46
TOTALS			208	498	427	368	82	63

Table 2: Use of some medical terms found in authoritive sources by oto-rhino-laryngeologists. See text for details.

In table 3, some details are shown related to the terms that the clinicians proposed themselves as alternatives for the terms found in the original corpus: “Np” is the number of clinicians that proposed their “preferred” term, “Nt” the number of terms that have been proposed by the clinicians, “Pmax” the number of clinicians that proposed the term that most often was proposed by all. This “overall preferred” term is shown in column 2.

From table 3, it can be calculated that - given that 28 terms were processed - 209/28 = 7.46 terms were proposed per original term, whereas over all cases, the “overall preferred” proposed terms received a mean support of 208/506 = 41%.

The same calculations were done for the 20 terms of table 2, yielding a mean of 4.95 proposals per term, and an overall mean support for the best proposals of 52%.

4. Discussion

Table 1 and table 2 (column (2)) show that the responding clinicians did not have problems in understanding the terms used by their colleagues [MRQ: a, b]. The one notable exception is “livide slijmvliezen” the meaning of which was not clear to 35% of the clinicians. Comparing table 1 and table 2, it might be possible to infer that the terms found in the literature (table 2) were generally understood by more clinicians than the terms found in the original corpus. Indeed, no term in table 2 scored under 96%.

Original term	Proposed term with highest support	Np	Nt	Pmax
aspecifieke mucosahyperreactiviteit	aspecifieke hyperreactiviteit vd mucosa	38	12	7
klassieke inhalatie-allergenen	aerogene allergenen	13	8	3
acute exacerbaties	acute opstoot	33	5	21
livide slijmvliezen	bleek slijmvlies	20	8	9
congestieve neusmucosa	gezwollen neusslijmvlies	12	6	4
hypertrofe onderste neusschelpen	hypertrofie van de onderste neusschelpen	18	6	6
habituele neusademhaling	gewone neusademhaling	8	5	3
habituele mondademhaling	habituele mondademing	10	3	6
nasale klank	nasaliteit	19	7	6
moeilijke neusademhaling	neusobstructie	26	10	14
etmoidale sluiering	ethmoidale sluier	6	3	4
etterige secreties	purulente secreties	10	4	6
obstructief ademen	belemmerde ademhaling	17	10	3
crypteuse amandelen	cryptische amandelen	19	10	10
kissing tonsils	hypertrofische amandelen	15	10	4
dysphagie voor vloeistoffen	slikstoornis voor vloeistof	3	3	1
dysfonie bij intensieve stembelasting	heesheid	17	10	3
onvolledige stembandsluiting	onvolledige sluiting van de stembanden	17	10	5
epitympanale trommelvliesretracties	atticale retractiepocket	19	12	4
auditieve communicatievaardigheid	spraakverstaanbaarheid	6	5	2
discreet transmissieverlies	geleidingsverlies	30	13	11
neurosensorieel verlies	perceptieverlies	24	5	16
echte draainissen	rotatoire vertigo	46	12	25
carotis souffle	carotis geruis	15	4	7
faciale pijnen	aangezichtspijnen	25	5	16
maxillaire tandpijn	maxillaire pijnen	18	13	4
opgezette halsbasis	zwelling van de halsbasis	17	8	4
koude nodulus	koude nodus	5	2	4
TOTALS		506	209	208

Table 3: Support for proposed alternative terms. See text for details.

On the other hand, though the terms were judged to be understandable, very few terms were generally indicated as being used “very often” or “always”. As a group, the terms found in the literature appear to be accepted somewhat better than the terms found in the original corpus (use-measure 63 versus 50), though whether this is significant at the level of individual terms is doubtful.

Very meaningful on the other hand are the data related to the “proposed preferred terms” given by the various respondents (table 3). We don’t exaggerate (too much) when we claim that each clinician has almost its own preferred term ! Some exceptions are “aangezichtspijnen” instead of “faciale pijnen”, “rotatoire vertigo” instead of “echte draainissen”, “acute opstoot” instead of “acute exacerbaties”, and “neusobstructie” instead of “moeilijke neusademhaling”. The fact that for 11 terms out of the 28 at least 10 alternatives were proposed, is striking [MRQ: c] !

There were (as a mean) less terms proposed for the terms found in the literature as compared to those coming from the original corpus [MRQ: d].

Although the results of this work seem to show that terms from the literature have a slight tendency to be better understood by clinicians, the actual use of them is not significantly more frequent. Given the MRQs listed below, a careful study design will probably lead to no difference at all.

The work indicates that clinicians use many synonymous expressions when reporting clinical findings in their patients. This variability is kept under control in electronic healthcare record systems that use controlled vocabularies to guide data entry. Controlled vocabularies may fit nicely in systems with graphical user interfaces, but are extremely difficult to use with speech recognition systems that accept voice input from the user and send the resulting text to natural language understanding applications for further analysis and structuring of the data [1]. Such systems must account for the various ways in which a clinician can say the same thing !

The high variability of expression in every day clinical language makes controlled vocabularies and even other medical nomenclatures of little use for automatic text understanding. In 1995, the UMLS contained 371.742 terms and 190.863 concepts which means a synonym ratio of 1.95 [5]. In 2000, the figures are respectively 1.338.650 versus 730.155, yielding a ratio of 1.83 [6]. SNOMED International proposes 26.312 synonyms amongst 128.855 terms, yielding a ratio of 1.26 [7].

This is much lower than the 7.46 and 4.95 ratio's found in this study, figures of this size also being reported in [8] where 12.180 (24.3 % of total registrations) free text entries used to indicate the chief complaint in an ER setting could be reduced to 3% by identifying synonymic expressions (ratio 8.0) and forcing a controlled vocabulary to be used.

5. Conclusion

This study suggests that clinicians do not face major problems in understanding terms derived from clinical narratives generated by peers. It also suggests that "preferred terms" are merely an academic artefact than a reality: claiming that there are as many preferred terms for a medical concept as there are physicians, is far from a witticism. As long as clinicians are kept chained in front of a screen, controlled vocabularies with preferred terms can be used to guide data entry. But as soon as speech recognition systems will dominate graphical user interfaces, their role will change. Clinicians will want (and get) back the freedom of expression with all delicate yet important nuances that are required for individual patient care. Natural language understanding applications will have to take over the responsibility to map free text to coded entries. In such systems, preferred terms will not be a source for data entry, but a target for language understanding [9].

6. References

[1] 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.

[2] Van der Meer E, Hoffmann J (eds) Knowledge aided information processing. North-Holland, Amsterdam - New York - Oxford - Tokyo, 1987.

[3] 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.

[4] Moerkerke C. Sintomatología en ORL: estudio terminográfico y análisis crítico. Afstudeerscriptie in de Vertaalkunde, 1999-2000, Mercator Hogeschool Gent.

[5] Tuttle MS, Suarez-Munist ON, Olson NE, Sherertz DD, Sperzel WD, Erlbaum MS, Fuller LF, Hole WT, Nelson SJ, Cole WG, Lipow SS. Merging Terminologies. In Greenes RA, Peterson HE, Protti DJ (eds.) Proceedings of MEDINFO'95, 1995, 162-166.

[6] Unified Medical Language System, 11th edition, January 2000. National Library of Medicine.

[7] Zweigenbaum P, Grabar N. A contribution of Medical Terminology to Medical Language Processing Resources: Experiments in Morphological Knowledge Acquisition from Thesauri. Proceedings of IMIA-WG6, Medical Concept and Language Representation, Phoenix, 16-19/12/1999, 155 - 167.

[8] Aronsky D, Merkley K, Haug P, James B. Reducing free text entries: a continuous quality improvement project. In Chute CG (ed.) Proceedings of AMIA '98, 1998: 969.

[9] Ceusters W. Language, medical terminologies and structured electronic patient records: how to escape the Bermuda triangle ? (in press)

APPENDIX: Methodology Related Questions

[a] Do the respondents constitute a representative sample for all ORL-clinicians, or did the clinicians that have problems in understanding the terms preferentially not return the questionnaire ?

[b] Can we trust that the clinicians that claimed to understand the terms, understood them actually in exactly the same sense ?

[c] Not all respondents gave an alternative for all terms. If no alternative is given, it is not clear whether this is due to the known "open-question reluctance" phenomenon or whether the original term was judged to be good enough.

[d] Though in the questionnaire no explicit difference was made between terms from the source corpus or the literature, the terms coming from the corpus were listed first. It would have been better to mix both groups randomly. Now the question may rise whether or not the lower number of proposed alternatives for the literature-group is the result of the well-known "fatigue-effect" for open questions at the end of a questionnaire.