Nissilä, N. & N. Siponkoski (toim.) 2012. Kielet liikkeessä, Språk i rörelse, Languages in Motion, Sprachen in Bewegung. VAKKI-symposiumi XXXII 11.–12.2.2012. VAKKI Publications 1. Vaasa, 118–129.

International Medical Classifications from the Terminological Point of View (cases: International Classification of Diseases and Terminologia Anatomica)

Igor Kudashev
Palmenia Centre for Continuing Education
University of Helsinki

Terminologinen työ on tiettävästi työlästä, ja terminologisten tietojen uudelleenkäyttö säästäisi paljon aikaa ja vaivaa. Kerron tässä artikkelissa haasteista, joihin olemme törmänneet yrittäessämme digitalisoida, jäsentää ja ontologisoida kahta terveydenhuoltoalan kansainvälistä luokittelua kieliteknologiahankkeen tarpeisiin. Toinen luokitteluista on Maailman terveysjärjestö WHO:n ylläpitämä ICD-10 Tautiluokitus ja toinen on anatominen luokittelu Terminologia Anatomica. Näytän esimerkkien avulla, mitkä leksikografiset käytänteet heikentävät luokittelujen käyttäjäystävällisyyttä ja hankaloittavat niiden käyttöä terminologisten tietojen lähteenä mm. kieliteknologiasovelluksissa. Päädyn siihen johtopäätökseen, että hakuteosten laatua ja käyttäjäystävällisyyttä voi parantaa ja niiden käyttöalaa laajentaa, jos niitä suunnitellaan ja laaditaan alusta asti yhteistyössä terminologien ja kieliteknologian asiantuntijoiden kanssa. Pohdin myös teoreettisia kysymyksiä, joita luokittelujen käsittely on herättänyt, mm. luokitteluissa käytettyjen luokkien nimien ja tunnisteiden luonnetta ja asemaa.

Keywords: terminology work, terminology management, classification, medical classification

1 Introduction

Reusability of data is one of central desiderata in language processing and terminology management. In this paper, I describe some challenges that we experienced during the ontologization of two international medical classifications for the purposes of research projects aimed at creating new NLP (natural language processing) applications.

The first classification is the *International Statistical Classification of Diseases and Related Health Problems, 10th Revision* (commonly referred to as "ICD-10") maintained by the World Health Organization (WHO), and the second one is *Terminologia Anatomica*, which is the international standard on human anatomic terminology developed by the International Federation of Associations of Anatomists (IFAA) and the Federative Committee on Anatomical Terminology (FCAT).

I am going to analyze the two classifications from the terminographical point of view and make some proposals as how the user-friendliness and reusability of such classifications could be improved in future editions, with particular focus on the needs of terminology work and natural language processing. I also draw the readers' attention to some theoretical questions that arose during the processing of the above mentioned classifications.

2 Overview of the research projects

In this section, I provide some background information about the research projects in which the digitalization and ontologization of ICD-10 and TA were required. In the *ContentFactory* project (2007–2011), a prototype of the *TermFactory* platform was created that allows storing terminological and lexical data in the form of ontologies, so that they can be used not only by human users but also by various natural language processing applications (see Kudashev, Carlson & Kudasheva 2010 for a detailed description of the platform).

The platform is currently being tested in several other research and development projects, such as an FP7 machine translation project *MOLTO*[1] (Multilingual Online Translation) and a Tekes project *Mobster*[2] (Mobiili ja ympäristöön integroitu sanelu- ja kommunikointisovellus terveydenhuoltoalalle). The latter project aims at providing reliable and user-friendly dictating and communicating applications for medical personnel.

3 Overview of the medical classifications

International Statistical Classification of Diseases and Related Health Problems, 10th Revision ("ICD-10") contains codes of diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or morbidity. ICD-10 has been translated from English into 42 languages and is used in over one hundred countries, including Finland (see Sosiaali- ja terveydenhuollon luokitukset 2012), for cause of death reporting and statistics, in clinical and research work, social insurance, etc.

ICD-10 contains about 14 400 predefined codes. This number can also be expanded up to 16 000 with the help of optional codes and even further with codes

[1] http://www.molto-project.eu
[2] http://www.helsinki.fi/palmenia/kouvola/2011/27012011.htm

that are meant to be reported in separate fields. The latest electronic version of ICD-10 is available on WHO's website[3].

Terminologia Anatomica was published as a book in 1998, and its scanned version is available on the website of the International Federation of Associations of Anatomists[4]. The classification contains terminology for about 7 500 human macroscopic anatomical structures. The main language is Latin, and English equivalents are provided for reference. *Terminologia Anatomica* has been translated in many languages. In Finland, it was the main source of Latin anatomical terms in the *Medical Dictionary* (Lääketieteen termit 2007) published by the Finnish Medical Society *Duodecim*.

4 Multipurposeness of reference products as a tendency and desideratum

As Hartmann has pointed out in his book 'Teaching and Researching Lexicography' (2001: 5), the boundary lines between lexicographic and non-lexicographic reference products are disappearing fast under the impact of computer technology. Hartmann has also predicted that eventually all these might come under the umbrella of a yet to be developed 'reference science'.

Our research projects on machine translation and speech recognition support at least the first observation. To my mind, it is very desirable that reference products which are rich in terminology should be developed from the very start in close co-operation with professional terminologists, terminographers and computer linguists. On the one hand, such co-operation will enhance the quality and user-friendliness of reference products, and on the other hand it will improve reusability of data, save efforts and prevent errors during the extraction of terminological information. Next, I will formulate some principles of a 'terminologist-friendly' reference work using some negative examples from ICD-10 and TA.

5 Requirements related to terms

Terms in classifications often require transformations before they can be extracted. This means additional work, often manual, and also increases the risk of errors and misinterpretations. In the ideal case, terms should be provided according to 'context-free' orthographic and grammatical rules. A mnemonic formula that

[3] http://www.who.int/classifications/icd
[4] http://www.unifr.ch/ifaa/Public/EntryPage/ViewSource.html

is applicable here is FUN: the form of the terms should be Full, Unambiguous, and Natural. In the subsections to follow, I provide examples of problems related to the form of terms in classifications.

5.1 Omissions

A major problem in using classifications and other reference products as sources of terminology is omission of parts of terms. For example, in TA many terms of the second and lower hierarchical levels are provided in their abridged form (e.g. *Deep nodes* instead of *Deep popliteal lymph nodes (of lower limb)* – see *Example 1*). The reconstruction of the full form of a term can not be done automatically. Besides, it requires excellent command of the LSP in question. In such cases, the benefits of space-saving are definitely outweighed by the risk of mistakes and information loss.

(1) A13.3.05.001 Lymph nodes of lower limb
A13.3.05.011 Popliteal nodes
A13.3.05.013 Deep nodes

In ICD-10, one can also find incomplete 'contextual' names of classes like *Tongue, unspecified* (*Example 2*) while the full name of the class would be *Malignant neoplasm of tongue, unspecified*.

(2) C02 Malignant neoplasm of other and unspecified parts of tongue
…
C02.9 Tongue, unspecified

5.2 Capitalization

Only those elements of terms may be capitalized that would normally be capitalized in the middle of a sentence. Otherwise capitalization of initial or all letters should not be used. Some words in terms may contain capital letters (e.g. *HIV disease resulting in Kaposi's sarcoma*), so mechanical bringing of all terms to the lower case is not an option. In ICD-10 and TA initial letters of terms are capitalized.

5.3 Abbreviations

Standard abbreviations (such as *HIV* – human immunodeficiency virus) and short forms may be provided as synonyms but the main term should always be written in full. Classification-specific abbreviations should not be used. These rules are not always followed in ICD-10.

5.4 Substitutions

Recurrent parts of terms are quite often substituted in classifications with a special mark, for example a dash. However, as *Example 3* from the Russian index of TA demonstrates, such presentation is not always user-friendly and is not suitable for inflective languages, as the canonical form of the term can not always be reverted automatically. In *Example 3*, if the dashes are substituted with the corresponding words, the resulting expression will not make sense, as the words need yet to be inflected and the word order needs to be changed.

(3) Желудочек левый A12.1.04.001 100
--клапан аорты A12.1.04.012 100
----заслонка венечная левая A12.1.04.014 100
------правая A12.1.04.013 100
-----невенечная A12.1.04.015 100

5.5 Inversions

In classifications, inverted order of term elements is frequently used. Inversion allows saving space and clustering of similar terms around the main noun. However, terminologists need to extract terms in their basic form, and often there is no way for them to know whether the inversion is classification-specific or whether the term is actually used in this way in the medical discourse. For example, what is the natural order of elements for the term in the bulleted list in *Example 4*? The most frequent and natural form is *familial hyperkalaemic periodic paralysis*, but a terminologist has to make some inquiries before he or she finds that out.

(4) Periodic paralysis (familial)
- hyperkalaemic

5.6 Inconsistency

The form of the term should be consistent throughout the classification, for example in different entries, in the body and in the indexes, etc. Otherwise terminologists may wonder whether different terms relate to the same thing or different and which one of the forms is the 'official' or the 'correct' one. For example, in ICD-10 *herpesviral meningitis* is given in two different forms in another entry: *meningitis herpesviral* and *meningitis due to herpex simplex*.

5.7 Clustering

Each term in classifications should stand alone and no clustering should be used. Variants and synonyms should be separated from the main term visually and/or with mark-up. In ICD-10, parentheses are used for optional and interchangeable elements and square brackets for synonyms. For example, according to the ICD's User's Guide, in the string *Hypertension (arterial)(benign)(essential)(malignant) (primary)(systematic)*, the word *Hypertension* may be used alone or when qualified by any, or any combination, of the words in parentheses. However, an attempt to automatically generate legitimate terms from a string with multiple optional components may fail. In *Example 5* from ICD-10, *vena cava* can not be simultaneously *inferior* and *superior*, so an automatically generated term *absence of vena cava inferior superior* will not make sense.

(5) Absence of vena cava (inferior)(superior)

Inclusion of synonyms in the main term may lead to mistakes. In *Example 6* from ICD-10, a layman can not tell for sure if *Hansen's disease* is a synonym of *leprosy* or *arthritis in leprosy*. This ambiguity has lead to a translation mistake in the Russian version of ICD-10.

(6) Arthritis in:
- leprosy [Hansen's disease] (A30.-+)

Clustering may also be less obvious. Special use of 'and' in place of 'and/or' in ICD-10 packs three terms into one. For example, according to the ICD's User's Guide, in the rubric *Tuberculosis of bones and joints*, are to be classified cases of 'tuberculosis of bones', 'tuberculosis of joints' and 'tuberculosis of bones and joints'. Such 'hidden clustering' is undesirable from the point of view of terminology extraction.

5.8 Foreign elements

Terms may not include any foreign elements, such as notes or parts of codes. For example, parts of codes, such as *(E10-E14 with common fourth character .3)*, *(C00-D48+)* in *Example 7*, have migrated into otherwise well-structured electronic versions of the Finnish and Russian translations of ICD-10 as parts of terms.

(7) H28.0* Diabetic cataract (E10-E14 with common fourth character .3)
M36.1* Arthropathy in neoplastic disease (C00-D48+)

5.9 Ambiguous character of data and type of relations

Different types of data should be marked unambiguously in classifications, and the relations between them should be made explicit. For example, synonyms of terms are scattered in ICD-10 between a) the main term where they are sometimes given in square brackets, b) so called *inclusion terms* which are provided just below the main term, and c) the Alphabetical Index. Inclusion terms and index terms are provided only as a guide to the rubrics, so their relations with the main terms remain unclear. They may be full synonyms, partial synonyms, related terms, special cases, etc. This prevents their effective use alongside the main terms.

6 Requirements related to concept identification

Most classifications are concept-oriented, and concepts have identification numbers. From the point of view of terminology management, these IDs are as important as terms, so there are certain requirements and recommendations related to the form of concept IDs as well.

The spelling conventions of codes should be consistent throughout the classification as well as its translations and localizations. Spelling conventions must be strict, binding and well-documented. Codes should not be split in parts, i.e. no foreign elements should be allowed within the code (such as terms, notes, etc.), like in *Example 8* from ICD-10.

(8) G73.1* Eaton-Lambert syndrome (C00-D48+)

Codes should not contain any special characters which may be difficult to replicate or which may cause other technical problems. Meanwhile, ICD-10 uses the dagger symbol (†) and TA the Mars/Venus symbol (♂ / ♀) as parts of codes (*Example 9*).

(9) A08.3.01.021 M. vesicoprostaticus ♂
A08.3.01.021 M. vesicovaginalis ♀

7 Requirements related to availability and usability of classifications

In order to be machine-processable, data should be available in machine-readable and modular form. However, for example, TA is available only in the printed form and a as poor quality image scan.

Authenticity and correctness of data are vital for the use of classifications as sources of terminological information. This is why official versions of classifications should be available from the official site of organizations that are responsible for their maintenance. The main official site should also provide links to the official translations of the classification. For example, the authenticity of the electronic versions of the Russian translation of ICD-10 was questionable, which is why it had to be proofread against the official paper version.

Electronic versions of classifications should also be complete. For example, the plain text version of ICD-10 does not contain inclusion and exclusion terms, and the index is not available in the electronic format either, at least from the official site.

The format of classifications should be rich enough in order to capture all essential features of the classification, such as its structure, cross-references, and inline formatting – italics, upper and lower indexes, etc. On the other hand, different projects and target groups have different needs, so simpler formats, such as comma-separated values, or Excel table, should also be available if possible. Whatever formats are used for presenting a classification, they should be well-documented. Version management, too, should be made easy, for example with the help of conversion tables and Upgrade Manuals.

As the needs of human readers and machine agents vary, classifications should be available in both human- and machine-oriented formats. As machine-oriented formats are stricter and less ambiguous, they should be created in the first place, and human-readable views should be generated automatically from them. This is one of the visions and research goals of the *ContentFactory* project and its spin-offs.

8 Theoretical questions

In this section, I would like to discuss some theoretical questions that arose during the processing of ICD-10 and TA.

8.1 Are names of classes in ICD-10 terms?

According to a relatively fresh ISO definition (ISO 704:2009: 34), term is a *designation* consisting of one or more words representing a *general concept* in a *special language* in a specific *subject field*. *General concept* is defined in ISO 1087-1:2000 as 'concept which corresponds to two or more objects which form a group by reason of common properties', and *concept* as 'unit of knowledge created by a unique combination of characteristics'. Let us check some names of ICD-10 classes against these definitions.

All names of classes in ICD-10 are designations consisting of one or several words used in a specific subject field. Do they represent general concepts? Let us have a look at some classes from ICD-10 (*Example 10*):

(10) S43 Dislocation, sprain and strain of joints and ligaments of shoulder girdle
K57.5 Diverticular disease of both small and large intestine without perforation or abscess
B57.4+G05.2 Encephalitis, myelitis or encephalomyelitis in Chagas' disease
V86 Occupant of special all-terrain or other motor vehicle designed primarily for off-road use, injured in transport accident
D53.1 Other megaloblastic anaemias, not elsewhere classified

The nature of the concepts that stand behind the names of these classes is somewhat specific. Some names of classes (such as D53.1 in *Example 10*) contain deictic elements like 'other', 'not otherwise specified', 'not elsewhere classified' which imply that the corresponding concepts are context-bound. In practice they do not make sense outside the scope of the classification.

Other classes are complex, i.e. a few separate concepts are combined in them. For example, *encephalitis* and *myelitis* are individual diseases, and *encephalomyelitis* is their combination. However, in many ICD-10 classes they are put together and sometimes also combined with other diseases, as in class B57.4+G05.2 in *Example 10* above.

In spite of their special character, these concepts correspond to the definition of a general concept, as they denote groups of objects grouped by common properties. But are the designations of such concepts a part of the special (medical) language? The answer to this question depends on the chosen perspective. If we focus on the manifestations of the language (texts), we will probably locate these designations in abundance. If we ask whether these designations are a part of the conceptual space and lexis in a given national medical LSP, the answer would probably be 'no'.

There are several reasons for the negative answer. First, the function of ICD-10 classes is limited and very pragmatic: they help keep statistics of diseases and causes of death. They almost totally lack the cognitive function that is characteristic of scientific concepts and terminology. Their intention and designations are 'frozen' and do not evolve until the next revision of the classification. They are relevant only as long as the correspondent classification is valid. Second, ICD-10 can not compete (at least for the time being) with national concept systems and terminology. It aims at mapping national concepts and terms into international classes rather than substituting them. Third, the complexity, length and occasional clumsiness of some names of ICD-10 classes prevent them from becoming lexical units in medical LSP.

So, even if some names of ICD-10 classes formally coincide with medical terms, this should be understood as homonymy between two different systems and areas of application: dynamic, multifunctional medical LSP and a static, single-functional classification. Besides, some names of ICD-10 classes are of special character and do not have counterparts in living medical terminology.

8.2 Do names of classes in ICD-10 belong in a term bank and in which capacity?

Another question I would like to discuss is whether names of ICD-10 classes belong in a terminological reference product, for example in a national term bank. The most probable type of information that users of a term bank would

presumably seek about the names of ICD-10 classes is their foreign equivalents and relations with national medical terminology. The same information is valuable to NLP applications.

In terminological reference resources like term banks, terminological descriptions provide information on the objects of description (lexical LSP units) that is supposed to help users understand, use or substitute these objects in external contexts (Kudashev 2007). The range of units that may be included in a term bank is not restricted to terms. For example, term banks may contain lexicalized LSP units – instructions, commands and other set phrases (Kudashev 2010).

Although some of the names of ICD-10 classes coincide with terms and some other resemble lexicalized LSP units, we would recommend keeping classifications like ICD-10 as a separate, read-only collection that is only linked to a term bank but not merged with it. Term banks and classifications have different functions and different conventions that should not interfere with each other. At the same time, classification can be a valuable source of terminological information and can complement the terminological description provided in the term bank.

8.3 Do codes of classes belong in a term bank and in which capacity?

Means of formal notation (such as codes, formula, catalogue names, international scientific names) are often used in LSP texts interchangeably with the corresponding terms, and sometimes they are the only existing designation of a special object or concept. If there is a parallel lexical designation, means of formal notation usually become a part of the terminological description (Kudashev 2010).

All ICD-10 and TA codes have correspondent verbal designations, so they can be included in term entries as a part of terminological description, for example in the section devoted to describing the relations of a given term. As they are interchangeable in certain contexts, they could probably be treated as cross-code, context-bound synonyms. The range of contexts in which this relation is valid, should be strictly and explicitly specified by a reference to the name and version of the classification.

9 Conclusion

The main points of this article can be summarized as follows.

- Term banks and classifications have different goals and conventions and should be stored and managed as separate though interlinked resources.
- At the same time, classifications and other non-lexicographic reference products rich in terminology are valuable sources of terminological information, such as terms, definitions and concept relations.
- From the point of view of cost-effectiveness, it is very desirable that classifications should be designed from the very start as multipurpose and multiuser products and include natural language processing applications and other machine agents as their 'target group'.
- This goal can be achieved only if reference products rich in terminology are developed in close co-operation with professional terminologists, terminographers and computer linguists.

References

Hartmann, Reinhard Rudolf Karl (2001). *Teaching and Researching Lexicography*. New York: Longman.

ISO 704:2009 Terminology Work – Principles and Methods. Geneva: ISO.

ISO 1087-1:2000 Terminolgy Work – Vocabulary – Part 1: Theory and Application. Geneva: ISO.

Kudashev, Igor (2007). *Proektirovanie perevodčeskih slovarej special'noj leksiki* [Designing LSP Dictionaries for Translators]. Helsinki University Translation Studies Monographs 3. Helsinki: Helsinki University Print.

Kudashev, Igor (2010). What can be an Object of Terminological Description in a Term Bank? In: *Kieli ja tunteet. Käännösteoria, ammattikielet ja monikielisyys. VAKKI-juhlasymposiumi XXX. Vaasa 12. –13.2.2010,* 142–152. Eds. Niina Nissilä & Nestori Siponkoski. Vaasa: Vaasan yliopisto.

Kudashev, Igor, Lauri Carlson & Irina Kudasheva (2010). TermFactory: Collaborative Editing of Term Ontologies. In: *Terminology and Knowledge Engineering Conference 2010. Presenting Terminology and Knowledge Engineering Recourses Online: Models and Challenges*, 479–500. Eds. Úna Bhreathnach & Fionnuala Barra-Cusack. Dublin: Fiontar, Dublin City University.

Lääketieteen termit (2007). *Duodecimin selittävä suursanakirja*. 5. uudistettu painos. Helsinki: Duodecim.

Sosiaali- ja terveydenhuollon luokitukset (2012) [online]. Terveyden ja hyvinvoinnit laitos. [cited 30.4.2012]. Available at: http://www.thl.fi/fi_FI/web/fi/tutkimus/palvelut/ koodistopalvelu/esittely/luokitukset.