International Journal of Medical Informatics (2006) **75**, 542—552

ELSEVIER

International Journal of Medical Informatics

www.intl.elsevierhealth.com/journals/ijmi

Amplification of Terminologia anatomica by French language terms using Latin terms matching algorithm: A prototype for other language

Paul Fabry[a,*], **Robert Baud**[b], **Anita Burgun**[c], **Christian Lovis**[b]

[a] *CRED, Centre Hospitalier Universitaire de Sherbrooke, Qué., Canada*
[b] *Service d'Informatique Médicale, Hôpital Universitaire de Genève, Switzerland*
[c] *Laboratoire d'Informatique Médicale, Faculté de Médecine, Rennes, France*

Received 29 December 2004; received in revised form 23 August 2005; accepted 24 August 2005

KEYWORDS
Anatomy;
Terminologia anatomica;
Nomina anatomica;
Foundational model of anatomy;
Terminology mapping;
French language

Summary
Objective: Terminologia anatomica is the new standard in anatomical terminology. This terminology is available only in Latin and English and its worldwide adoption is subject to the addition of terms from others languages. On the other hand, *Nomina anatomica*, the previous standard, has been widely translated. Aim of this work was to append foreign terms to *Terminologia* by using similarity-matching algorithm between its Latin terms and those from *Nomina*.
Methods: A semi-automatic matching of Latin terms from *Terminologia* with those of *Nomina* was performed using a string-to-string distance algorithm and manual assessment. We used a French—Latin version of *Nomina* together with *Terminologia* and we suggested French terms for *Terminologia*. Coverage was evaluated by the number of exact and approximate matches. A target of 78% was set due to the higher number of terms in *Terminologia* compared to *Nomina*. Relevance was estimated by manually comparing the meanings of the English and French terms related to the same Latin term. The question was whether they refer to the same anatomical structure.
Results: Exact or approximate matches were found for 5982 terms (76.5%) of *Terminologia*. Our results indicated that more than 75% of the terms from *Terminologia* came from *Nomina*, most of them were left unchanged and all were used with the same meaning.
Conclusion: This method produces relevant results, reaching our 78% target. The method is based only on Latin terms and can be used for other languages. We consider this work as a starting point for adding terms to other knowledge sources, such as the foundational model of anatomy or the Unified Medical Language System (UMLS).

* Corresponding author. Tel.: +1 819 346 1110x12710; fax: +1 819 820 6853.
E-mail address: paul.fabry@usherbrooke.ca (P. Fabry).

1386-5056/$ — see front matter
doi:10.1016/j.ijmedinf.2005.08.008

1. Introduction

Human anatomy is a fundamental medical science and it is of interest for most domains of medicine. Individualizing the parts and the structures of the human body is a process that took centuries and occurred in many countries, each using its own language. It was estimated that there were some 50,000 anatomical terms in use at the end of the 19th century but they applied only to some 5000—6000 structures [1]. Not only several terms were related to the same structure (synonyms) but one term could be used to name different structures (homonyms) depending on its author. In addition, many of these terms were eponyms, i.e. terms including the name of a person.

In order to overcome this obstacle, some anatomists called for a standard vocabulary in anatomy, i.e. the enumeration of the anatomical structures under a common language, a universally agreed name and without ambiguity.

After several attempts (*Basler Nomina anatomica* in 1895, *Jenensa Nomina anatomica* in 1939), the International Federation of Associations of Anatomists (IFAA) led to the publication of the first international terminology of anatomy, the *Parisiense Nomina anatomica*, known as *Nomina anatomica*, in 1955 [2]. *Nomina* included a list of 5640 terms in Latin to be translated into existing languages, pertaining to macroscopic anatomy.

From 1955 to 1989, *Nomina* had six editions with additions in embryology (*Nomina embryologica*) and histology (*Nomina histologica*) but with only minor modifications in anatomy compared to the original version. Meanwhile, the advances in medicine, particularly in surgical procedures and neurosciences, led to the individualization of new anatomical structures, and therefore, to the creation of new anatomical terms, especially in regard to the central nervous system.

In order to provide standardization for these new terms, and to overcome emerging disagreements about *Nomina* [3], the IFAA created the Federative Committee on Anatomical Terminology (FCAT) in 1989, which succeeded with the publication of *Terminologia anatomica* in 1998 [4]. *Terminologia* provides a list of some 7500 structures relative to macroscopic anatomy. In addition to Latin terms, *Terminologia* includes their English counterparts in current usage and the most frequent eponyms, and indexes all the terms using alphanumerical codes.

Terminologia has been approved by the IFAA and is now considered as the new international standard in anatomical terminology. However, to support its diffusion in non-English-speaking countries, important work is needed to translate its terms and to match them with the multiple synonyms and eponyms present in the anatomical language. On the other hand, *Nomina* is still widely used as a reference throughout the world and its Latin terms are associated with terms in other language, such as French, Spanish, etc.

An approach for easing the translation of *Terminologia* is to map this terminology with existing sources for multilingual terminologies and anatomy, such as the Unified Medical Language System (UMLS) and the foundational model of anatomy [5]. Indeed, several authors have already evaluated the UMLS for integrating others terminologies [6,7]. However, as it is exposed below, while a significant number of *Terminologia* terms can be linked to UMLS concepts, associating these terms with terms in foreign languages give uneven results.

Another approach is to use a bilingual version of *Nomina* and match its Latin terms with those of *Terminologia*. If we consider the terms at their string level, in other words as plain sequences of characters, this match can be done automatically using computerized tools.

In this article, we present a methodology using a string-to-string distance algorithm to perform a semi-automatic matching of Latin terms from *Terminologia* with those of *Nomina*. The method was evaluated by using a French—Latin version of *Nomina* together with *Terminologia* in order to suggest French terms for *Terminologia*.

2. Materials and methods

In the beginning of this section, we describe the characteristics of the two terminologies, Nomina anatomica and Terminologia anatomica. Then, we give an account of others sources of anatomical knowledge, such as UMLS and the FMA, before reporting the methodology.

2.1. Anatomical terminologies

2.1.1. Nomina anatomica

Nomina provides a list of approximately 5800 Latin terms for its last edition. The terms are not associated with definitions of the structures to which they refer, though footnotes explain why some terms have been chosen. The order of terms follows a classification by functions and regions into nine chapters: general terms, regions and body parts, osteology, arthrology, myology, splanchnology, angiology, nervous system and sense organs. In addition, text indentations are used to represent hierarchical relationships between the terms (Fig. 1).

Arteria auricularis posterior
A. stylomastoidea
A. tympanica posterior
Rami mastoidei
†(Ramus stapedialis)
Ramus auricularis
Ramus occipitalis

Fig. 1 The posterior auricular artery and its branches in Nomina anatomica, fourth ed. †The term in brackets refers to an inconstant anatomical structure.

In order to limit the size of *Nomina*, a compositional approach was adopted. For example, in Fig. 1, text indentations designate the term ''*Rami mastoidei*'' (mastoid branches) as a child (or hyponym) of ''*A. tympanica posterior*'' (posterior tympanic artery). The term ''*Rami mastoidei*'' alone is not meaningful and the user has to infer that the proper term is ''*Rami mastoidei arteriae tympanicae posterioris*'' (mastoid branches of the posterior tympanic artery). In addition, the type of the hierarchical relationship between these two terms is not explicitly defined. A human reader deduces easily that mastoid branches are branches (i.e. offshoots that derive from a larger source) of the posterior tympanic artery and not a *kind of* posterior tympanic artery, or a *part of* this artery. For computer programs, on the other hand, a description of the type of relationships is required to make use of all the knowledge implied in these relationships.

This example displays other characteristics of the terms in *Nomina*: the presence of abbreviations (A. for *arteria*), and the morphological variants due to the Latin declensions (*arteria* in nominative, *arteriae* in genitive singular or nominative plural).

There is no bilingual version of Nomina directly available but its terms are usually found associated with terms in other languages in various anatomy textbooks. Moreover, anatomical term lists in languages other than Latin or English have not been generated.

2.1.2. Terminologia anatomica

Terminologia share common features with the fifth edition of Nomina but it also departs from it in several aspects. *Terminologia* also uses a classification by function and regions, though more precise as it includes sixteen chapters. *Terminologia* keeps the use of Latin and includes the terms of *Nomina* with minor differences. For example, the lymph nodes are designated as ''*Nodi lymphatici*'' in *Nomina* and as ''*Nodi lymphoidei*'' in *Terminologia*.

Terminologia includes also the same type of abbreviations as in *Nomina*. In addition, some terms include additional information within square brackets (Nervus glossopharyngeus [IX], for example) or special characters (♀ and ♂) to identify gender specific structures.

The compositional approach is maintained and the hierarchical relationships between terms are represented using text indentations and styles of headings, without improving the explicitness of the type of relationship in *Nomina*.

Terminologia, however, provides English terms. Eponyms are also present but in an appendix. In addition, *Terminologia* has a more complex structure: each anatomical structure is represented by a code, which is used to index all the terms in Latin or English related to this structure (Fig. 2).

The presence of codes allows one to overcome term ambiguities and synonymy. In addition, these codes may be used to index terms in other languages to create a multilingual version.

Terminologia provides a list of some 7500 codes associated with some 8400 Latin terms, 8500 English terms and 700 eponyms. There are approximately 1500 more anatomical structures in *Terminologia* than in Nomina, related to the central nervous system for the most part.

2.1.3. Knowledge representation in anatomical terminologies

An important point is that *Nomina* and *Terminologia* are fundamentally textbooks designed for human readers. The functional classification, the

A12.2.05.037	**Arteria auricularis posterior**	**Posterior auricular artery**
A12.2.05.038	A. stylomastoidea	Stylomastoid artery
A12.2.05.039	A. tympanica posterior	Posterior tympanic artery
A12.2.05.040	Rr. mastoidei	Mastoid branch
A12.2.05.041	†(R. stapedius)	(Stapedial branch)
A12.2.05.042	R. auricularis	Auricular branch
A12.2.05.043	R. occipitalis	Occipital branch
A12.2.05.044	R. parotideus	Parotid branch

Fig. 2 The posterior auricular artery and its branches in Terminologia anatomica. A new structure, Ramus parotideus (parotid branch) has been individualized compared to Nomina anatomica.

text indentations and the compositional approach provide an intuitive way for readers to navigate through the terms. However, from a knowledge representation point-of-view, several characteristics of these terminologies may prevent to use them as a formal representation of anatomy.

As no definition is provided, in some cases it is difficult to identify the structure, which is designated by the term. The duodenum, for example, is classically divided into four parts: superior, descending, inferior, and ascending. *Terminologia* adds a fifth: ''*Hidden part of duodenum*'' (**A05.6.02.010**), at the same hierarchical level than the four others. This term refers to the segment of the superior and descending parts of the duodenum underneath (hidden by) the liver but most anatomy textbooks do not name it separately.[1] On the other hand, reaching consensus on a definition for each structure referenced by a term is an enormous task. An approach could be to use anatomical drawings where terms are visually linked with the related structures.

The text indentations do not specify the type of relationships. This is a problem since an indentation may imply several different meanings. We have seen the ''Branch of'' relationship in the example above, but more than 80% of the relationships in *Terminologia* are ''part of'' and a few are ''kind of''. Each of these relations has distinct properties and the type of relationship connecting two terms has to be identified.

The classification of anatomical structures by their functions or localization is well known by those who study anatomy. However, this organization leads to many ambiguities and do not provide the required consistency for a formal model of the human body [8]. On the contrary, in a formal model, an anatomical structure should be associated with multiple properties considering its functions or localization. In addition, a formal model provides a method for generating consistent relationships and definitions [9]. Nevertheless, it should be understood that such formal approaches do not contradict *Terminologia*, but they complement it.

The objective of these terminologies is to provide a standard vocabulary in anatomy, and consequently to associate a unique term with each structure. Yet, as numerous terms are currently used, it may be useful to relate as many terms as possible to these structures [10]. In addition to the inclusion of synonyms, linguistic information could be added, such as plural forms or declensions, especially for Latin.

[1] The definition of the ''hidden part of the duodenum'' was given to the author by a member of the FCAT.

2.2. The foundational model of anatomy and the UMLS

The foundational model of anatomy (FMA)[2] [December 22, 2003 Version] is the reference ontology in the domain of anatomy and provides formal definitions and relationships of detailed anatomical concepts in a format ''understandable'' by computer programs [5]. The FMA includes four parts among them two are pertinent:

- The anatomy taxonomy (AT), which classifies the anatomical structures according to their physical properties. The AT encompasses all the anatomical structures in a network of ''kind of'' relationships. For example, long bones and flat bones are ''kind of'' bones, which mean that they share properties with their common ancestor (*genus*), but have additional properties which distinguish one from the other (*differentiae*). While both are bones, a long bone is different from a flat bone.
- The anatomical structural abstraction (ASA), which includes several taxonomies representing a particular type of constraint or relationship for anatomical structures. For instance, the ASA includes a ''dimensional ontology'', which assigns a geometrical shape to each structure, and a ''part of'' network, which represents anatomical structures as parts of one another. These networks allow organizing the anatomical structures according to different views.

The FMA provides more than 70,000 concepts and 110,000 terms, and also includes 7316 *Terminologia* codes as of this writing.

The Unified Medical Language System provides a formal framework for unifying various medical vocabularies [11,12]. One of its knowledge sources, the metathesaurus, links the terms having the same meaning to one unique concept.

The latest version of the metathesaurus as of this writing (2004AB) brings together 113 medical vocabularies in 17 languages for a total of 1,078,246 concepts and 3,866,024 terms.[3] The prominent languages in the metathesaurus are English (55.2% of the terms) and Spanish (30.4% of the terms). French is in sixth position with 1.2% of the terms.

Neither *Terminologia* nor *Nomina* are present in the metathesaurus. It includes, however, an important source of anatomical terms by means of the University of Washington Digital Anatomist (UWDA) vocabulary [13]. The UWDA includes 61,387 concepts and 90,752 terms and is a subset of the

[2] http://fma.biostr.washington.edu.

[3] What we consider as terms are the UMLS' normalized concept names (LUIs).

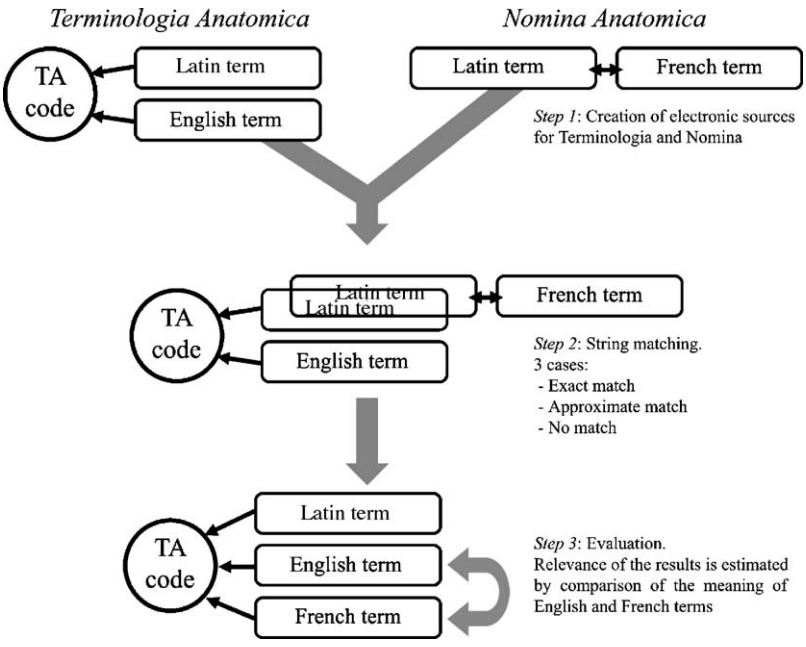

Fig. 3 Schematic view of the three steps of the methodology.

terms from the FMA. The UWDA, however, does not include *Terminologia* codes.

By performing an exact string match between the terms from the FMA and those of the UWDA, 6134 *Terminologia* codes (83.4% of the codes present in the FMA) were linked to an UMLS concept. Yet, using these concepts to relate UMLS terms in other languages to *Terminologia* codes gives uneven results. While 3839 codes (52.5% of the codes present in the FMA) were related to a Spanish term (thanks to the Spanish version of the SNOMED-CT included in the UMLS), lesser results were obtained for others languages: 663 codes (9.1%) for the Portuguese, 656 (9%) for the Dutch, 603 (8.2%) for the French and 589 (8.1%) for the German. These results are due to the fact that the only source of anatomical terms in these languages comes from the translations of the MeSH [14]. Therefore, a version of *Terminologia* in these languages requires another method.

2.3. Methodology

The main part of this work is based on approximate string matching, to be introduced below. Nomina and Terminologia were considered as raw lists of terms, disregarding all the knowledge implied by the position of a term in the list and its relationships with adjacent terms. In addition, we consider anatomical terms at their string level. Electronic versions of these terminologies, such as databases or text files are not directly available and had to be created. The methodology was designed in three main steps (Fig. 3):

- Creation of electronic sources for *Terminologia* and *Nomina*.
- Semi-automatic mapping of Latin terms including a pre-processing step, a string-to-string distance computation and a manual validation. An automatic validation was also tested.
- Evaluation of the coverage and the relevance of the results by comparing English and French terms.

These steps are detailed below.

2.3.1. Creation of electronic sources

For *Nomina*, we used the index of a French anatomy textbook [15] and a French anatomical dictionary [16]. We chose these sources instead of the original *Nomina* hardcopy for two reasons: first, they provide French terms related to those of *Nomina*; second, they do not have a compositional approach as in *Nomina* and the Latin terms are fully extended. These sources were processed with an optical character recognition (OCR) software.[4] We manually corrected the errors generated by the OCR and created a list including couples of Latin—French term.

The task was easier for *Terminologia* as its codes and Latin terms (extended) are present in the FMA.

[4] In order to respect the intellectual property rights, these electronic sources are considered as private and will not be distributed by the authors.

Nevertheless, we also processed the hardcopy of *Terminologia* with OCR software to generate our own version and combined it with the one present in the FMA to be as exhaustive as possible.

2.3.2. String matching

As minor differences exist between Latin terms of *Nomina* and *Terminologia*, an exact string matching may generate too much silence in the results. An algorithm was developed using edit distance for matching approximate strings.

Prior to the distance computation, all terms were pre-processed for normalization. This pre-processing consisted of:

- conversion to all lower case;
- removal of punctuation signs and brackets (for *Terminologia*, special characters and information enclosed within square brackets were also removed);
- expansion of abbreviations;
- alphabetical sorting of the words in each term.

All the pre-processing steps were performed automatically. While automatic expansion of abbreviation has proven to be a difficult task [17,18], the abbreviations present in our corpuses were unambiguous and easily expandable (a. for arteria, nn. for nervi, etc.). Morphological variants due to Latin declensions were left unchanged.

Then, for each normalized term of *Terminologia*, the edit distance was measured with all the normalized terms from *Nomina*.

Edit distance is a well-known approach for measuring the distance between two strings [19]. It defines a set of edit operations on characters, such as deletion, insertion or substitution, together with a cost for each operation. Given two strings *x* and *y*, the distance between them is the minimum number of edit operations required to transform *x* into *y*. We used a rather simple edit distance known as the Levenshtein–Damereau distance [20,21], which assigns the same cost for each edit operation (Fig. 4).

Thus, in Fig. 4, the distance between ''*rami*'' and ''*ramus*'' is 2, because a substitution and an insertion are sufficient to transform the string ''*rami*'' into the string ''*ramus*'', and the cost associated with these operations is 2 (as defined in Fig. 4). For two equivalent strings, the distance is null.

Three cases were considered for string matching:

- One of the strings from *Nomina* has a null distance, meaning that the strings are equivalent: there is an exact match.

		r	a	m	u	s
	0	1	2	3	4	5
r	1	0	1	2	3	4
a	2	1	0	1	2	3
m	3	2	1	0	1	2
i	4	3	2	1	1	2

Three different string operators are defined: substitute, delete and insert. In the above diagram, values associated with each of these operations are the following:

Substitute (a,a) = 0
Substitue (a,b) =1
Delete (a) = *Insert (a)* = 1

Fig. 4 Computation of the Levenshtein distance between ramus and rami. The distance is the number in the bottom-right cell.

- All the strings from *Nomina* have a distance superior to 0. Then, the five closest strings are provided for manual validation.[5] The strings provided for validation are the original strings, prior to the normalization process. If one of these strings is validated: there is an approximate match.
- Finally, if none of the five closest strings are validated: there is no match.

This method requires significant manpower for terms validation. In order to decrease the human intervention, we also tested an alternative method based on the automatic selection of the closest term. However, automatically matching all term from *Terminologia* with the closest term from *Nomina* results in an important number of incorrect matches. So, we defined a threshold and exclude all matching terms having a distance value superior to this threshold.

For the string-matching task, we developed a software in DELPHI© allowing us to compute the distance between two lists of strings and providing to the user the five closest string for validation.

2.3.3. Evaluation

The former steps resulted in the generation of a list of quadruplets including a code from *Terminologia*, the corresponding Latin and English terms, and a French term.

Three points were examined for the evaluation:

- Coverage of the results: an evaluation of the coverage was given by the number of *Terminologia* codes found in the list. However, as there are

[5] This number of five has been experimentally chosen after a few tests on the method.

Table 1 Number of matches found for terms and codes from Terminologia, with the percentages and the average distance value

	Number of strings	Strings (%)	Number of TA codes	TA codes (%)
Exact matches	4863	57.6	4579	61.8
Approximate matches	1119	13.2	1092	14.7
Total matches	5982	70.8	5671	76.5
No match	2467	29.2	1744	23.5
Total	8449	100.0	7415	100.0

approximately 1700 more anatomical structures in *Terminologia* than in *Nomina*, the expected coverage could not exceed 78%.
- Relevance of the results: the meanings of English and French terms related to the same code were manually compared by a physician. The question was whether they refer to the same anatomical structure.
- Automatic validation: we evaluated this alternative method with different thresholds.

3. Results

3.1. Creation of electronic sources

For *Nomina*, a list of 13,407 couples Latin—French term was generated. This list included numerous synonyms among the French terms related to the same Latin term. Thus, the list was separated in two:

- a first list composed of the different Latin terms associated with a unique French term (8885 couples);
- a second list for the synonyms including 4522 couples French—French term synonym.

The number of Latin terms obtained in our corpus (8885) is surprisingly large compared to the size of Nomina (about 5800 terms). This is due to the presence of variants among the Latin terms and to the fact that the used sources included also Latin terms from *Nomina histologica* and *Nomina embryologica*. The mean number (±S.D.) of words per terms was 2.5 (±1) and the mean number (±S.D.) of characters per strings was 22.2 (±9.5).

Regarding *Terminologia*, we found 7418 different codes associated with 8450 Latin terms and 9204 English terms (including the eponyms). However, two codes were found associated only with an English term (**A05.6.03.003**: *Meckel's diverticulum* and **A06.6.02.018**: *Sibson's fascia*)[6] and were removed. In addition, the code at the top level of the hierarchy (**A00.0.00.000**: *Anatomia generalis, General anatomy*) was also removed.

The resulting list was composed of 7415 codes related to 8449 Latin terms and 9201 English terms. Considering the Latin terms, the mean number (±S.D.) of words per terms was 3.1 (±1.4) and the mean number (±S.D.) of characters per strings was 26.8 (±12.2).

3.2. String matching

The 8449 Latin terms from *Terminologia* were compared with our list of 8885 Latin terms from *Nomina*. Exact or approximate strings matches were found for 5982 terms of *Terminologia*, related to 5671 codes (Table 1).

The ratio between exact matches and total matches is important and represents 83.1%. The large majority of Latin terms from *Nomina* present in *Terminologia* have been left unchanged.

3.3. Evaluation

The methodology provided a French term for 5671 (76.5%) codes of *Terminologia*. An excerpt of the results is given in Table 2. The overall coverage is satisfactory taking into account the limit of 78% due to the higher number of terms in *Terminologia*. In addition, the coverage was evaluated for each chapter of Terminologia (Table 3).

The chapter with the lowest coverage (64.2%) deals with the nervous system. This is not surprising as this is the chapter that includes most of the additional terms compared to Nomina.

The relevance of the results was evaluated by comparing the meanings of French and English terms related to the same code from *Terminologia*. This comparison was done manually with the help of medical dictionaries [22] and anatomical atlases [23,24].

[6] This code (**A06.6.02.018**) may be the consequence of a syntax error in *Terminologia* as Sibson's fascia is the eponym for ''suprapleural membrane'' (membrana suprapleuralis) associated with the code **A07.1.02.018**.

Table 2 Results for the posterior auricular artery and its branches

TA code	Latin term	English term	French term
A12.2.05.037	Arteria auricularis posterior	Posterior auricular artery	Artère auriculaire postérieure
A12.2.05.038	Arteria stylomastoidea	Stylomastoid artery	Artère stylo-mastoïdienne
A12.2.05.039	Arteria tympanica posterior	Posterior tympanic artery	Artère tympanique postérieure
A12.2.05.040	Ramus mastoidei arteriae tympanicae posterioris	Mastoid branch of posterior tympanic artery	Branche mastoïdienne de l'A. tympanique postérieure
A12.2.05.041	Ramus stapedius arteriae tympanicae posterioris	Stapedial branch of posterior tympanic artery	?
A12.2.05.042	Ramus auricularis arteriae auricularis posterioris	Auricular branch of posterior auricular artery	Branche auriculaire de l'A. auriculaire postérieure
A12.2.05.043	Ramus occipitalis arteriae auricularis posterioris	Occipital branch of posterior auricular artery	Branche occipitale de l'A. auriculaire postérieure
A12.2.05.044	Ramus parotideus arteriae auricularis posterioris	Parotid branch of posterior auricular artery	?

The ''?'' indicates that no match have been found.

Table 3 Coverage for each chapter of Terminologia expressed in number and percentage of matched codes

Chapter	Title	Number of codes	Number of matches	%
A01	General terms	259	206	79.5
A02	Bone system	981	891	90.8
A03	Articular system	384	329	85.7
A04	Muscular system	646	517	80.0
A05	Digestive system	490	365	74.5
A06	Respiratory system	248	221	89.1
A07	Thoracic cavities	27	21	77.8
A08	Urinary system	100	70	70.0
A09	Genital system	299	241	80.6
A10	Abdominal and pelvic cavities	93	81	87.1
A11	Endocrins glands	34	27	79.4
A12	Cardiovascular system	1275	898	70.4
A13	Lymphatic system	202	134	66.3
A14	Nervous system	1808	1161	64.2
A15	Sense organs	506	454	89.7
A16	Teguments	63	55	87.3
Total		7415	5671	76.5

No difference was found between them. The terms from *Nomina* present in *Terminologia* are used with the same meaning.

The strings selected in approximate matches are sorted by their ranks in Table 4. More than 76.5% of the matching strings were the strings with the smallest distance. On the other hand, there is no correlation between the distance value and the rank.

Two main categories of modifications were identified between our corpus of terms from Nomina and the terms from Terminologia.Modifications due to the use of different words: we have already cited the example of ''*lymphoideus*/*lymphaticus*'' and there are others, such as ''*posterior*/*dorsalis*'', ''*anterior*/*ventralis*'', etc.Modifications due to the

Table 4 Number and percentage of approximate strings matches sorted by their rank

Rank	Matches	%	Distance value (mean ± S.D.)
1	858	76.7	8.3 ± 6.1
2	155	13.9	15.3 ± 6.5
3	46	4.1	15.6 ± 6.7
4	40	3.6	17.2 ± 8.1
5	20	1.8	19.3 ± 8.7
Total	1119	100.00	

Table 5 Number of total matches and correct matches found for Terminologia using the selection of the closest term and a threshold

Threshold	Total matches	Correct matches	Precision (%)	Recall (%)
5	530	290	54.7	25.9
10	1747	622	35.6	55.6
15	2500	727	29.1	65.0
20	3060	820	26.8	73.3
25	3335	847	25.4	75.7
30	3483	857	24.6	76.6

extension of the terms: because of the compositional approach taken by both *Nomina* and *Terminologia*, the terms require to be extended and variants may appear during this process. For example, the term ''*Lamina horizontalis*'' (horizontal plate) need to be extended using its hypernym ''*Os palatinum*'' (palatine bone). Three different extensions were found in Nomina and Terminologia:

- ''*Lamina horizontalis* (*Os palatinum*)'': hypernym with brackets.
- ''*Lamina horizontalis ossis palatini*'': declension of the hypernym (genitive).
- ''*Lamina horizontalis palatini*'': declension and truncation of the hypernym.

English and French terms also showed these kinds of modifications.

In addition, no approximate match due to an OCR error was found.

We removed all the strings having an exact match from *Terminologia* and performed the automatic validation on the 3586 remaining strings. The results of manual validation were used as a gold standard to test this alternative method. Recall and precision were estimated for each threshold. Precision was estimated as the ratio between the number of correct matches and the total number of matches. Recall was calculated as the ratio between the number of correct matches and the total number of correct matches found by manual validation (1119). The results are given in Table 5.

Using a small threshold gives the best precision (54.7%). On the other hand, we lost about 800 matches compared to the manual validation. Recall increases significantly when using higher thresholds though it cannot exceed 76.7% because 23.3% of the strings manually validated were not the closest strings. In addition, high thresholds give low precision: the automatic validation requires also important manual assessment to eliminate incorrect matches in the result.

4. Discussion

Terminologia anatomica constitutes a major improvement of *Nomina anatomica*. More than just adding terms, *Terminologia* provides a semantic structure for representing the anatomical knowledge in human-readable form. We distinguish three parts in *Terminologia*: an enumeration of the objects constituting the human body (designated by codes), a list of the terms associated with these objects, and a representation of the relationships between these objects. *Terminologia* has been sanctioned by an international group of anatomists and should be considered as a new standard in anatomical terminology. However, as this terminology is only available in Latin and English, its worldwide adoption is subdued to the addition of terms from others languages.

A translation of the terms from *Terminologia* is a labour-intensive task, requiring a lot of manpower from domain experts. Our goal was to propose a practical method to append terms in others languages to the anatomical structures enumerated in *Terminologia*.

Several approaches have already been explored to align anatomical terminologies [25—27]. However, the problems we faced were distinct from those works, in that one of our sources existed only in hard copy with no explicit relationship between the terms. Using structural methods to map the two terminologies would have required to manually rebuild the semantic structure of *Nomina*. Instead, our work focused only on the terms and did not take into account the knowledge implied by the semantic structure.

The expected similarity of the Latin terms present in the two terminologies was an important motivation for this work. Indeed, *Terminologia* had a conservative approach towards the terms of *Nomina* despite the evolution of the structure. Our results outlined several points about the place of *Nomina* within *Terminologia*: more than 75% of the terms from *Terminologia* came from *Nomina*, most

of them were left unchanged and all were used with the same meaning.

We used a rather basic string-to-string edit distance, compared to the numerous distances, which have been developed [21]. However, these distances usually require large strings or additional processing, such as part-of-speech tagging to be fully efficient. Token-based distances, such as the Dice coefficient, when applied for the comparison of phrases, consider terms as ''bags'' of words and use the number of equivalent words between two terms to compute a distance. In a preliminary study, we tested these distances with poor results due to the low number of words by terms (2.5 ± 1 for *Nomina* and 3.1 ± 1.4 for *Terminologia*). On the other hand, the Levenshtein—Damerau distances do not need prior knowledge about the terms and has relevant results on relatively small strings. In addition, this distance has already been used with significant results for spell checking in medical texts [28].

Evaluation shows good results with an overall coverage reaching the 78% limit. However, an important limitation of this work was the lack of validated source in an electronic format. These sources had to be created and we have no assurance of exhaustiveness. For example, we identified 7418 codes, 8450 Latin terms and 9204 English terms in *Terminologia* but the textbook includes no references about the total number of codes or terms and we cannot be sure to have created an exhaustive version.

More important is the compositional approach taken by both terminologies. The variation in terms due to their different extensions was the main cause of approximate matches and lowered accuracy of automatic validation.

French terms were used as it is the native language of the authors. Yet, this work's strength is that only the Latin terms were involved in the matching process. Therefore, the method is suitable for any other languages, to the condition that they have their own version of *Nomina*. The FCAT confirmed the Latin as the language of the *definitive terminology* [3] because of its neutrality and international character. In addition to this role of ''depository'' of the anatomical knowledge, we suggest that Latin may still take an active part in the creation of multilingual terminologies.

The knowledge implied in the semantic structure of *Terminologia* is accessible for humans but this structure lacks the requirement to be relevantly processed by computer programs [8]. Ontology, by providing a formal definition of the concepts of a given domain and by representing all the relationships between these concepts, is a recognized method for expressing biomedical knowledge in a computer accessible format. The reference ontology in anatomy is the FMA. As it includes terms and codes from *Terminologia*, a multilingual version of *Terminologia* can be used to add terms from other languages to the FMA.

Finally, another step towards a multilingual *Terminologia* could be reached by adding *Terminologia* to the UMLS metathesaurus, whether as a specific source or by expanding the UWDA vocabulary.

5. Conclusion

In this work, we tested a method using the similarity between Latin terms from *Nomina anatomica* and *Terminologia anatomica* to expand the latter with French terms. This method allows minimizing human resources with relevant results, reaching a specified target of 78% due to the fact that Terminologia includes 1700 more terms than *Nomina*. In addition, as the method is based only on Latin terms, it could be used for other language as well. However, further manual work and validations by experts are yet to be done to produce a comprehensive translation of *Terminologia*. Finally, we consider this work as a starting point for adding terms to other knowledge sources, such as the FMA or UMLS.

Acknowledgements

This work has been funded by the Swiss National Science Foundation (SNF 632-066041). The authors wish to thank Dr. Henning Müller for his help.

Summary points

What is already known on this subject

Terminologia anatomica is considered as the new international standard in anatomical terminology. However, it provides only Latin and English terms and much work is needed to translate its terms in foreign languages and to match them with the multiple synonyms and eponyms present in the anatomical language

Terminologia anatomica includes the Latin terms from the previous standard *Nomina anatomica* with minor differences

Nomina anatomica is still widely used as a reference throughout the world and its Latin terms are associated with terms in other languages, such as French, Spanish, etc.

What this study adds

By using a string-to-string edit distance algorithm, Latin terms form *Terminologia anatomica* can be semi-automatically matched with those of *Nomina anatomica* with relevant results

This matching process allows us to append French terms to the English terms included in *Terminologia anatomica*

Our study suggests that this method could be used for any foreign version of *Nomina anatomica*

References

[1] C. Rosse, P. Gaddum-Rosse, W.H. Hollinshead, Hollinshead's Textbook of Anatomy, fifth ed., Lippincott-Raven Publishers, Philadelphia, PA, 1997.

[2] International Anatomical Nomenclature Committee, Nomina anatomica, fifth ed, Approved by the Eleventh International Congress of Anatomists at Mexico City, 1980, Together with Nomina Histologica, second ed., and Nomina Embryologica, second ed., Williams & Wilkins, Baltimore, MD, 1983.

[3] I. Whitmore, Terminologia anatomica: new terminology for the new anatomist, Anat. Rec. 257 (2) (1999) 50—53.

[4] Federative Committee on Anatomical Terminology, Terminologia anatomica: international anatomical terminology, Thieme, Stuttgart, New York, 1998.

[5] C. Rosse, J.L. Mejino Jr., A reference ontology for biomedical informatics: the foundational model of anatomy, J. Biomed. Inform. 36 (6) (2003) 478—500.

[6] J.J. Cimino, S.B. Johnson, P. Peng, A. Aguirre, From ICD9-CM to MeSH using the UMLS: a how-to guide, Proc. Annu. Symp. Comput. Appl. Med. Care (1993) 730—734.

[7] Q. Zeng, J.J. Cimino, Mapping medical vocabularies to the Unified Medical Language System, Proc. AMIA Annu. Fall Symp. (1996) 105—109.

[8] C. Rosse, Terminologia anatomica: considered from the perspective of next-generation knowledge sources, Clin. Anat. 14 (2) (2001) 120—133.

[9] J. Michael, J.L. Mejino Jr., C. Rosse, The role of definitions in biomedical concept representation, Proc. AMIA Symp. (2001) 463—467.

[10] R.H. Baud, C. Lovis, A.M. Rassinoux, P. Ruch, A. Geissbuhler, Controlling the vocabulary for anatomy, Proc. AMIA Symp. (2002) 26—30.

[11] D.A. Lindberg, B.L. Humphreys, A.T. McCray, The Unified Medical Language System, Methods Inform. Med. 32 (4) (1993) 281—291.

[12] K.E. Campbell, D.E. Oliver, K.A. Spackman, E.H. Shortliffe, Representing thoughts, words, and things in the UMLS, J. Am. Med. Inform Assoc. 5 (5) (1998) 421—431.

[13] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Res. 32 (Database issue) (2004) D267—D270.

[14] UMLS AB Documentation, National Library of Medicine, Bethesda, 2004.

[15] H. Rouvière, Anatomie Humaine, Descriptive, Topographique et Fonctionnelle, 14e éd. revisée ed., Masson, Paris, 1997.

[16] P. Kamina, Petit dictionnaire d'anatomie, d'embryologie et d'histologie (Nomina Anatomica), Maloine, Paris, 1990.

[17] H. Yu, G. Hripcsak, C. Friedman, Mapping abbreviations to full forms in biomedical articles, J. Am. Med. Inform. Assoc. 9 (3) (2002) 262—272.

[18] P. Ruch, R. Baud, A. Geissbuhler, Evaluating and reducing the effect of data corruption when applying bag of words approaches to medical records, Int. J. Med. Inform. 67 (1—3) (2002) 75—83.

[19] C. Lovis, R.H. Baud, Fast exact string pattern-matching algorithms adapted to the characteristics of the medical language, J. Am. Med. Inform. Assoc. 7 (4) (2000) 378—391.

[20] F. Damereau, A technique for computer detection and correction of spelling errors, Commun. ACM 7 (3) (1964) 171—176.

[21] C.D. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, MA, 1999.

[22] WA.N. Dorland, Dorland's Illustrated Medical Dictionary, 30th ed., W.B. Saunders, Philadelphia, PA, London, 2003.

[23] F.H. Netter, J.T. Hansen, Atlas of Human Anatomy, third ed., Icon Learning Systems, Teterboro, NJ, 2003.

[24] J. Sobotta, R. Putz, R. Pabst, R. Putz, A.H. Weiglein, Sobotta Atlas of Human Anatomy, 13th English ed., Lippincott Williams & Wilkins, Philadelphia, 2001.

[25] P. Mork, R. Pottinger, P.A. Bernstein, Challenges in precisely aligning models of human anatomy using generic schema matching, Medinfo 2004 (2004) 401—405.

[26] K.L. Rickard, J.L. Mejino Jr., R.F. Martin, A.V. Agoncillo, C. Rosse, Problems and solutions with integrating terminologies into evolving knowledge bases, Medinfo 2004 (2004) 420—424.

[27] S. Zhang, O. Bodenreider, Aligning representations of anatomy using lexical and structural methods, AMIA Annu. Symp. Proc. (2003) 753—757.

[28] P. Ruch, R. Baud, A. Geissbuhler, Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record, Artif. Intell. Med. 29 (1—2) (2003) 169—184.