Ceusters W, Buekens F, De Moor G, Waagmeester A. The distinction between linguistic and conceptual semantics in medical terminology and its implications for NLP-based knowledge acquisition. Meth Inform Med 1998; 37:327-33.

The Distinction between Linguistic and Conceptual Semantics in Medical Terminology and its Implications for NLP-Based Knowledge Acquisition

Werner Ceusters [1, 2], Filip Buekens [3], Georges De Moor [2], Andra Waagmeester [4, 1]

[1] *Office Line Engineering NV, Zonnegem, Belgium*
[2] *Department of Medical Informatics, University Hospital Ghent, Belgium*
[3] *Department of Philosophy, University of Tilburg, The Netherlands*
[4] *Vakgroep Medische Informatiekunde, University of Amsterdam, The Netherlands*

Natural language understanding systems have to exploit various kinds of knowledge to be able to represent the meaning behind texts. Getting this knowledge in place is often such a huge enterprise that it is tempting to look for systems that can discover such knowledge automatically. In this paper, we describe how the distinction between conceptual and linguistic semantics probably can assist in reaching this objective, provided that distinguishing between them is not done to rigorously. We present several toy examples to support this view and argue that in a multilingual environment linguistic ontologies should be designed as interfaces between domain conceptualisations and linguistic knowledge bases.

Introduction

The purpose of the GALEN project is to develop language independent concept representation systems as the foundations for the next generation of multilingual coding systems [1]. At the heart of the project is the development of a reference model for medical concepts (CORE) supported by a formal language for medical concept representation (GRAIL) [2]. A particular characteristic of the approach is the clear separation of the pure conceptual knowledge from other types of knowledge, including linguistic knowledge [3], in order to arrive in the future to application-independent medical terminologies [4]. Although on a theoretical basis the feasibility of these objectives is debatable [5], actual work within the GALEN-IN-USE project shows that on a relatively concise domain such as surgical procedures, distributed collaborative modelling can be achieved over linguistic borders. As could be expected, the process is however extremely slow. Formal “naming” and subcategorisation of new concepts at the one hand, and (in)consistent modelling of natural language expressions using the building blocks of the CORE that already are available, turn out to be the most frequent reasons for discussion.

Given the very promising results of the MultiTALE-I semantic tagger for neurosurgical procedure reports [6, 7, 8], it was investigated whether or not this manual modelling work could be speeded up by tailoring the tagger for the automatic generation of GALEN-templates from natural language expressions out of the SNOMED procedure axis. These templates are a kind of intermediate representation used by the domain modellers in order not to be confronted with the complexity of the GRAIL language itself [9, 10]. This turned out to be feasible indeed, although a lot of efforts and resources had to be invested in providing sufficient medical knowledge to the parser for the delivery of acceptable results [11]. In fact, it became clear that contrary to what originally was expected, far more extra-linguistic knowledge was required to transform surgical procedure natural language expressions automatically into GALEN-templates with the expected level of detail. In addition, from the surface language of surgical procedure expressions alone, not enough conceptual knowledge could be derived to produce GALEN templates with a sufficient level of detail. As a result, the researchers working on the improvement of the MultiTALE-I tagger were in fact duplicating the work being done by the modellers. The question was then: what is actually the level of conceptual detail that in this particular domain can be discovered by a natural language processing system that exploits mainly linguistic knowledge and as little as possible domain-dependent knowledge ?

The purpose of this paper is not to give a definite answer to this question, but rather to indicate some directions in which future research has to be conducted.

Delineating the problem

In a broader context, our problem can be rephrased as such: *given a text in a particular language and in a specific domain, how much knowledge about that domain, and about the way that particular language is used to communicate in that domain, can be discovered by automatic processes that are linguistically* *driven, and that are both independent from the domain and language under scrutiny*. This is not the same goal as in traditional automatic text understanding where the exact meaning of utterances is to be discovered. The latter can be seen as the upper bound of the former. Another difference is that for text understanding quite a lot of a priori knowledge must be available, while for our purposes, we just want to know the amount of knowledge that is minimally required.

The different forms of knowledge that traditionally are claimed to be required for proper written text understanding are: *morphology*, *syntax*, *semantics*, *pragmatics* and *discourse* or *world knowledge* [12]. It is obvious that these forms of knowledge do not stand on their own, but that they are tightly related. At morphological level, inflection may be seen as a pure syntactic phenomenon, whereas compounding is merely guided by semantic principles. The actual form of a sentence depends amongst others on the situation under which a meaning is to be conveyed. As such pragmatics and discourse have an influence on syntax. Some authors even deny or reduce the distinction between some of these kinds of knowledge. Quine for instance showed that semantic knowledge and world knowledge cannot sharply be delineated [13].

In a machine learning perspective, when dealing with terminology rich domains, and with automated knowledge acquisition from written text understanding as a primary goal, it is possible to simplify the picture and to adopt a rather reductionist view. First, we can abstract away from the discourse level. Authors of medical textbooks, developers of terminologies or physicians writing patient reports, merely want to convey facts, and not to invoke emotions or to initiate actions by the reader. As such, we can limit our analysis to what in the speech-act literature is known as *constative inscriptions*, sentences uttered in a descriptive context [14], however without being too narrow as is the case in the traditional formal linguistic semantics scene where sentence-meaning is viewed as being exhausted by propositional content and is truth-conditionally explicable [15].

We also can abstract away from pragmatics - although not ignore its existence - as it is not our aim to provide theories on how context changes the surface forms of the expressions we are looking at. When looking to terminological phrases, we can certainly abstract away from indexical information. Terminological phrases by definition have to be self-explaining and do not refer to entities that are outside the domain covered.

In a monolingual environment, we could also ignore morphology, but as multilinguality is one of our main objectives, this would be too big a sacrifice. However, for the sake of simplicity and quietly assuming that the principles that govern word-formation are similar to the principles that govern syntax, we will not further deal with morphology in this paper.

Linguistic and conceptual knowledge

In our reductionist view, we can see a medical text, or more in particular the terms or rubrics of a medical terminological system such as SNOMED, as the product of a process in which words or word groups that refer to concepts, are put together following linguistic rules to form larger word groups that refer to new concepts that have a certain relationship with the original concepts.

Since the early activities of CEN/TC251, references to *conceptual* models, *concept* systems and *conceptual* semantics are dominating the medical informatics literature [16]. For the purpose of this paper, we mean by *conceptual knowledge* that knowledge that exclusively deals with concepts and the organisation of these concepts in a structure, independent of any language. This is not a fortiori the same as what in the linguistic literature is known as *conceptual semantics*, which is a particular theory on *meaning as conceptual structure* [17, 18]. Central in this theory is that semantic structures (what we denote) and conceptual structures (what we mean) converge, or even are the same. However, this probably is the case in a terminology rich domain such as medicine. Hence the *semantics* (i.e. the linguistic meaning) of a SNOMED expression can be said to be equal to the concept that is referred to.

In the light of our machine learning approach to written text understanding, we mean by *linguistic knowledge* that knowledge that specifies the rules of how valid expressions in a particular language are formed. This kind of knowledge comes in different flavours, two of which in our reductionist view are of importance. First there is the pure grammatical or syntactic knowledge that f.i. dictates phrase constituent order. Typical examples are the adjective - noun order in English, and the noun - adjective order in French. Gender agreement between nouns and adjectives in French is another example.

A second kind of linguistic knowledge is the one that is influenced by meaning. It is this kind of knowledge that tells us that actions usually are denoted by verbs, and entities by nouns. It is also this kind of linguistic knowledge that dictates us that adjectives denoting colour must appear just in front of nouns, and after other adjectives. This knowledge is extremely interesting for our purposes, as it holds the key of the door that leads from denotation to meaning. The particular branch of semantics that deals with this issue is *linguistic semantics*: *the study of literal meanings that are grammaticalised in a language* [19].

Linguistic semantics

A first principle of linguistic semantics is that one looks only at the *literal*, i.e. *decontextualised* meaning of an expression. From the standpoint of literal meaning, the expression

(E-1) *removal of cardiac pacemaker from epicardium or myocardium.*

represents a state of affairs that involves an event of *removing* and certain entities namely a *cardiac pacemaker*, an *epicardium* and a *myocardium*. There is no discussion about that. If we know that this expression is the rubric-term for SNOMED-code P1-315C4, then we know also that the implicational, i.e. contextualised meaning of this expression is that if on a patient a cardiac pacemaker is removed from one of the two specified places, this may be registered in his medical record as P1-315C4 if there is an agreement in the institution where the procedure is carried out that such interventions are coded in SNOMED. The notions *patient*, *institution*, *agreement*, etc, are required to understand the full semantics of the expression, but it is obvious that these notions are not encoded in the sentence itself. Hence they are not part of the linguistic semantics, or the *grammatical meaning* of the sentence.

At the other hand, the entities pacemaker, epicardium and myocardium appear in sentence (E-1) as structural categories, in casu nouns, that are essential to the formation of English sentences. From expression (E-1), we know also that it is the pacemaker that is the entity on which the event of removing acts, and not the epi- or myocardium. We are sure about that just because of the form of the expression in English, and not because we have to infer it from other information, e.g. because this expression is a rubric in SNOMED. It is the preposition *of* that marks the object that is removed, and the preposition *from* that encodes the source from which the removal is carried out.

The semiotic triangle revisited

Although meaning can be defined in several ways, the semiotic view is currently the most referred to in the medical informatics literature on natural language processing: meaning results of a relation between a *signifier* (an expression) and a *signified* (that to what is referred to) [20]. How this relation looks like, is a second point of debate, though mostly the conceptualist position (Figure 1) of Ogden and Richards is adhered to [21].

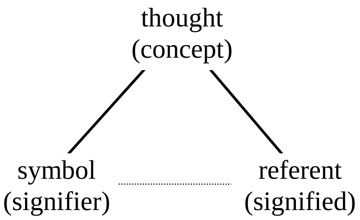


Figure 1 - The semiotic triangle.

For our purposes, we can generalise this triangle by taking into consideration:

1. all the signs that are used within the language(s) for which we want to develop natural language analysers,
2. all the concepts inside the terminological domain(s) that we want to cover,
3. the relationships that hold over the concepts mentioned in 2,
4. the rules that govern the allowed combinations of signs in the language under scrutiny, and,
5. the overt relationships between sign-combination and concept combination.

In doing so, we can reshape the triangle as in Figure 2.

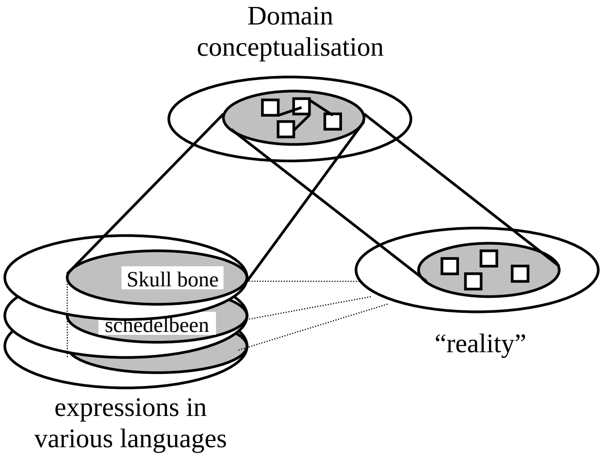


Figure 2 - The semiotic triangle from a computational linguistic semantics viewpoint

The problem revisited

Figure 2 can be used to understand our initial problem more intuitively. The hypothesis is that in medical texts, and especially in the rubrics of medical terminologies, "parts" of the conceptualisation of the domain is reflected in "parts" of the linguistic rules that dictate sentence formation. These "parts" are represented by the grey areas. It is not known what exactly the relative proportions of the grey areas are. Not inside a given language, and not between the areas at concept- and language level. Certain conceptual relations must dictate linguistic relations, and inversely, from the linguistic rules that possibly can be discovered in a language, some of them reflect conceptual relationships. Is it then possible:

1. to identify these relationships ?
2. to quantify the mutual influence ?
3. to design machine learning systems that are able to discover these various kinds of knowledge ?
4. to predict the results in terms of performance of such systems when varying the amount of knowledge (linguistic, conceptual, or both) they are entitled to start with ?

In the following paragraphs, we will present some toy examples in which a natural language analyser has to discover as much knowledge as possible. Deliberately, most of the examples given present an extremely oversimplified view of reality.

Discovering linguistic knowledge from scratch

As a starter, consider the following list of short phrases that are exemplary for the short telegraph style entries found in medical records.

big tumour	small lumb	malignant mass
small artery	dry skin	
tumour removal	skin incision	

Table 1 - Telegraph style medical phrases

Equipped with just the knowledge that each valid sentence has to be composed of a number of words separated by blanks, the only knowledge that can be deduced here is syntax: sentences of these language are of the form:

(E- 2) S → Word Word.

When studying the phrases more in detail, a subcategorisation of the words is possible just by looking to where they appear in the phrases. As such, three categories can be distinguished:

(E- 3) C1 = {big, small, malignant, dry}
(E- 4) C2 = {tumour, skin}
(E- 5) C3 = {artery, removal, incision, mass, lumb}

C1-words only appear in front position, C3-words only in tail position, and C2-words in both. As a consequence, the following syntactic rules can be derived:

(E- 6) S → C1 C2
(E- 7) S → C1 C3
(E- 8) S → C2 C3

Other types of analysis are possible. Well-known techniques are co-occurrence and distributional analyses for which numerous software packages exist. A particular application for such techniques is automatic thesaurus discovery or generation, which has been studied since the early years of automatic information retrieval, e.g. [22] and more recently [23]. The goal is different, but the principles behind the technique are the same.

The techniques are applied to investigate how particular words co-occur. Given our example above, and adding all possible combinations between words (with possible we mean syntactically and semantically acceptable although these notions do not necessarily have to be known by the algorithm that will analyse the phrases), we quickly notice that the words *big* and *small* fall into the same category. Also this kind of analysis will reveal that the nouns will be categorised differently than in (E- 4) and (E- 5), such that *removal* and *incision* will be separated from the others. Here the differentiating criteria is not the fact that words may appear in front- or tail-position, but whether they can take in front-position a word that itself may appear in tail-position in other constructions.

Combining these criteria, we finally then can come up with the following categories:

(E- 9) Ca = {big, small, malignant, dry}
(E- 10)Cb = {removal, incision}
(E-11) Cc = {artery, tumour, skin, mass, lumb}

Although without any additional information, a machine cannot induce why for other than pure distributional reasons the words are classified as above, humans quickly notice that category Ca is populated by adjectives, while categories Cb and Cc contain nouns. As such, one can say that our distributional induction algorithm discovered the existence of different parts of speech (word classes or lexical categories). Based on the examples, it then can build the grammar of the particular language under consideration:

(E- 12)S → Ca Cc
(E- 13)S → Cc Cb

Compared to the grammar of (E- 6), (E- 7), and (E- 8), relatively less non-sensical phrases can be generated, but at the same time, some acceptable phrases such as *small incision* can never be generated.

Discovering conceptual knowledge from scratch

At the same time, one can say that some semantic information is discovered. If a valid syntactic structure is formed according to (E- 12), the relation between Cc and Ca can be described as *HasFeature*, while in case of (E- 13) the relation between Cb and Cc would better be described as *ActsOn*. Of course, this is an information that with the current knowledge available, and with the small sample of phrases, only a human can obtain. The main point is here: there are characteristics in the surface language of the medical phrases presented, that reflect the semantics of what is being expressed.

Exploiting linguistic knowledge

If prior knowledge about parts of speech (POS) is available, then the language can be described by the following two syntactic rules:

(E- 14)S → Adj Noun
(E- 15)S → Noun Noun

These rules could be formulated by a linguist who studied the examples, but could also be induced automatically from the examples presented, given that for each word the POS is known.

Note that the grammar derived here is less powerful than the previous ones. Despite the availability of prior knowledge, more non-sensical combinations can be formed between the adjectives and nouns. In fact, when looking for correlations between sets of adjectives and sets of nouns, other syntactical patterns than just the adj - noun combination must be looked for. The *HasFeature* relation earlier mentioned can be realised in various syntactic environments, but it seems to be typical that in all these constructions the basic categories *noun* and *adjective* are connected either non-verbally as in (E- 12), or with a copula as in the following configurations:

(E- 16) Noun “is” Adjective
(E- 17) Noun “must be” Adjective
(E- 18) Noun “cannot be” Adjective

Not all adjectives can enter in such a configuration with all nouns, and further semantic information can be derived from that.

Exploiting multilingual knowledge

In a multilingual environment other sources of information can be used to discover such regularities. Table 2 shows a bilingual English - Dutch alienated list of words that are used in the (also alienated) phrases of Table 3. Without using any linguistic knowledge, an algorithm might discover that each of the phrases of Table 3 can be classified as belonging to one of two categories. In a first category, an acceptable phrase in one language can be produced by just substituting the words in the source language by their translations in the target language. In the second category, the order of the words has to be reversed as well. This is the case for “skin incision - insnede huid” and “tumour removal - verwijdering tumor”.

big	grote
small	kleine
tumour	tumor
removal	verwijdering
lumb	massa
skin	huid
incision	insnede
malignant	kwaadaardige
dry	droge
artery	arterie

Table 2 - Bilingual word-list (English - Dutch)

big tumour	grote tumor
small artery	kleine arterie
tumour removal	verwijdering tumor
small lumb	kleine massa
skin incision	insnede huid
malignant mass	kwaadaardige massa
dry skin	droge huid
small incision	kleine insnede

Table 3 - Bilingual phrase list (English - Dutch)

Based upon these observations, an algorithm might induce the special status of the words *incision* and *removal* as compared to *tumour*, *skin*, *mass*, *artery*, etc. As a consequence, similar conclusions can be drawn as previously.

Exploiting conceptual knowledge

Two possible situations can be thought of: one in which conceptual knowledge is overt available in the texts to be processed, and one where the natural language processor can use build-in conceptual knowledge to derive more knowledge.

Exploiting external conceptual knowledge

This technique can be used when parsing natural language expressions that appear in the rubrics of hierarchically organised classification systems such as SNOMED. Table 4 provides an example were there is a clear hierarchical relation between the terms, based on the concepts they represent.

P1-20000:	operative procedure on respiratory tract
P1-20000:	operation on respiratory tract
P1-21000:	operation on nose
P1-21100:	incision of nose
P1-21110:	conchotomy

Table 4 - Extraction from the SNOMED procedure axis, showing the concept related hierarchic organisation of terms.

A natural language processing system with only superficial semantic discriminating power such as MultiTALE-I can be used to augment lexicons semi-automatically, or to enrich its own conceptual knowledge. A simplified version of the basic MultiTALE-I lexicon that is needed to semantically tag the terms of Table 4 correctly, is represented in Table 5.

```
lex(“procedure”, ”noun_sg”, “deed”,
     [role(“anat”,”direct_object”)],
     [prep(“on”,”direct_object”)])
lex(“operative”, “adj”, “mod”,[],[])
lex(“respiratory tract”, “noun_sg”, “anatomy”,[],[])
lex(“operation”, “noun_sg”,“deed”,
     [role(“anat”,”direct_object”)],
     [prep(“on”,”direct_object”)])
lex(“nose”, “noun_sg”, “anatomy”, [],[])
lex(“incision”, “noun_sg”,“incise”,
     [role(“anat”,”direct_object”)],
     [prep(“of”,”direct_object”)])
```

```
lex(“conchotomy”, “noun_sg”, “incise”,
     [role(“#concha”,”direct_object”)],[])
```

Table 5 - Simplified MultiTALE-I lexicon required to tag the terms of Table 4.

This would result in the following tags for each term respectively:

```
deed(“operative procedure”) direct_object(“respiratory tract”)
deed(“operation”) direct_object(“respiratory tract”)
deed(“operation”) direct_object(“nose”)
deed(“incision”) direct_object(“nose”)
deed(“conchotomy”)
```

Table 6 - Semantic tagging result of MultiTALE-I on the rubrics of Table 4 using the lexicon of Table 5.

Based on the templates formed, all terms but the last having the pattern deed(X) direct_object(Y), the last one just deed(X), additional lexical information can be derived by taking advantage of the internal structure of SNOMED. According to the coding convention within SNOMED, we know that the first two expressions in Table 4 are synonyms, while each next term is “narrower than” the previous one. Hence, “operative procedure” must be a synonym of “operation”, “nose” must be a narrower term than “respiratory tract”, and “incision” must be “narrower” than ”operation”.

For the last expression, things are slightly more difficult. If we know for sure that *conchotomy* is narrower than *incision of nose*, then also the notion of direct_object must implicitly be represented in the term, and we can say that deed(“conchotomy”) is synonymous to deed(X)direct_object(Y) in which case one of the three following possibilities can be inferred:

1. X is more specific than the implicit deed in “conchotomy”, while Y is synonymous to the implicit direct_object, or
2. Y is more specific than the implicit direct_object, X being synonymous to the implicit deed, or
3. both X and Y are more specific.

Looking to the simplified lexicon in Table 5, we already know that X stands for “incise”, hence according to the second possibility, “concha” must be narrower than “nose”.

Again, we don’t claim here that this approach is immediately feasible as such. The example given has been carefully selected as it is well known that many vocabulary systems, thesauri and classifications are not that rigorously built, mixing different kinds of relationships in an unpredictable way [e.g. 24, 25]. This already is the case in the example given as the “narrower than” relationship between “incision” and “operation” refers to a *kind-of* relationship, while the one between “concha” and “nose”, and “nose” and “respiratory tract” refers to a *part-of* relationship.

Exploiting internal conceptual knowledge

Semi-automatic population of medical lexicons can also be achieved by using conceptual knowledge that is already available inside a natural language analyser. This possibility was exploited when upgrading MultiTALE-I to MultiTALE-II to meet the requirements of the GALEN project [11]. MultiTALE uses a kind of categorial grammar [26], and parsing proceeds in a bottom-up approach with syntactic and semantic constraints given equal importance as to the criteria upon which phrase constituents should be combined. MultiTALE-II is an improvement of MultiTALE-I in that phrase constituents with unknown syntactic or semantic categories can be given category labels according to internal constraints. This can be seen from the following examples where processing the SNOMED expression *“Removal of foreign body of iris by incision”* produces two results. The word “iris” is not encoded in the lexicon and two possible meanings are attributed to it: body_part or body_region.

```
RUBRIC "Removal of foreign body of iris by incision"
MAIN removal
        THEM_OBJ__IS foreign_body
                  LOCATION__IS *body_part
        TECHNIQUE__IS incision

RUBRIC "Removal of foreign body of iris by incision"
MAIN removal
        THEM_OBJ__IS foreign_body
                  LOCATION__IS *body_region
        TECHNIQUE__IS incision
```

Figure 3 - Two semantic parse results of the sentence “removal of foreign body of iris by incision”, the word “iris” not being coded in the lexicon of Multi-TALE II.

The syntactic parse of the expression reveals that also the part of speech tag “noun” has correctly been assigned to the word “iris”. It has to be noted that in Multi-TALE syntactic and semantic parsing occur in parallel, intermediate results at both levels being used for pruning the parse tree. For this reason, although the sentence in figure 1 and 2 is structurally ambiguous, only one syntactic parse is actually found in the output.

```
np      { { Removal of { foreign body of iris } } by incision }
np        { Removal of { foreign body of iris } }
noun         Removal
prep         of
np           { foreign body of iris }
noun           foreign body
prep           of
noun           *iris
```

prep	by
noun	incision

Figure 4 - Syntactic parse of the sentence “removal of foreign body of iris by incision”, the word “iris” not being coded in the lexicon of MultiTALE-II.

The mechanism for guessing categorial structures is not restricted to constituents that depend from a head constituent, but works also on head constituents themselves as can be seen in Figure 5 and Figure 6.

```
RUBRIC "xyz of drug"
MAIN *injection
        ACTS_ON drug
```

Figure 5 - GALEN template of the expression “*xyz of drug*”, produced by MultiTALE-II.

```
RUBRIC "xyz of bone"
MAIN *injection
        HAS_DESTINATION bone

RUBRIC "xyz of bone"
MAIN *surg_proc
        ACTS_ON bone
```

Figure 6 - GALEN templates of the expression “xyz of bone”, produced by MultiTALE-II.

The various possibilities with respect to the meaning of xyz and the associated semantic links, are derived from the internal concept hierarchy of MultiTALE-II, the relevant part of which being shown in Figure 7.

Figure 7 - Relevant part of the concept hierarchy of MultiTALE-II with respect to the expressions in Figure 5 and Figure 6.

From conceptual ontologies to linguistic ontologies

All knowledge based approaches rely on an *ontology*, a more or less formal 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 [27], though the need for an ontology in natural language processing applications is generally well accepted [28]. This is not to say that knowledge structuring based on a linguistic approach leads to the same result as when opting for a conceptual approach. A typical example is the ontological distinction between *nominal* and *natural kinds* [29], that in no language is grammaticalised just because the difference is pure definitional [30]. This again does not mean that such distinctions are not useful in a natural language processing applications. In MultiTALE-II for instance is the distinction between natural and nominal kinds used to analyse correctly expressions such as “capsulotomy of wrist” (Figure 8).

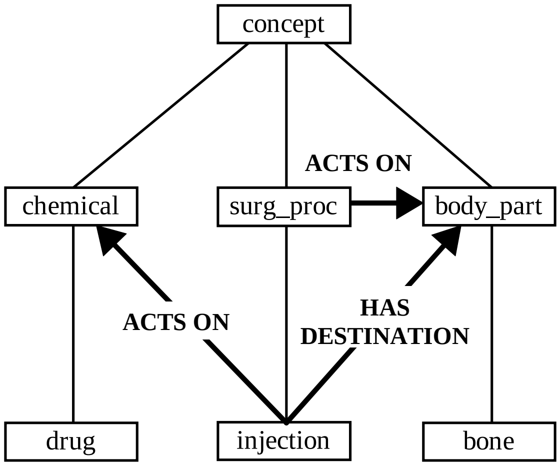
```
RUBRIC "capsulotomy of wrist"
MAIN capsulotomy
        ACTS_ON capsule
                HAS_LOCATION wrist
```

Figure 8 - GALEN template of the expression “*capsulotomy of wrist*”, produced by MultiTALE-II.

Situated ontologies - i.e. ontologies that are developed for solving particular problems in knowledge based applications [31] - that have to operate in natural language processing applications, are better suited to assist language understanding when the concepts and relationships they are built upon, are linguistically motivated [32].

In the perspective of re-usability, two dimensions have however to be explored: (relative) independence from particular languages and (relative) independence from particular domains.

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 [19]. 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 an ontology. Keeping the two independence criteria earlier mentioned in mind, a second revision of the semiotic triangle is needed. There clearly is a need for an additional layer between the sign-level and the concept-level as is outlined in Figure 9. In addition, at concept-level, there is not only one conceptualised



domain, but various conceptualisations have to be taken into account.

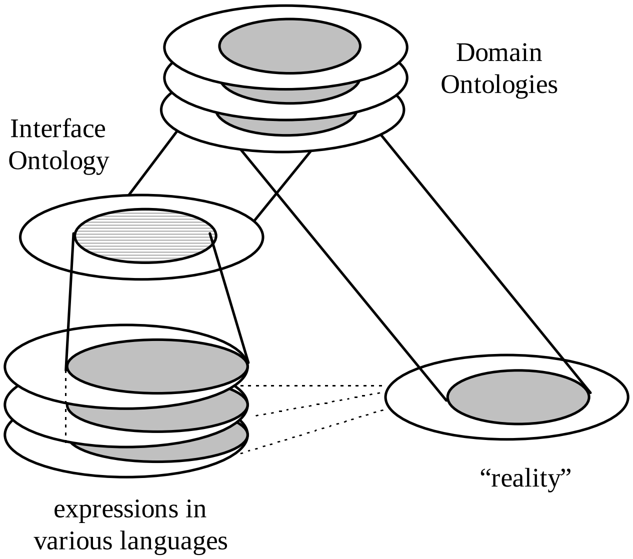


Figure 9 - The place of interface ontologies in the revised semiotic triangle.

The interface ontology approach starts to emerge in the computational linguistic literature. Approaches differ in the "distance" between the interface ontology and the domain ontologies at the one hand, and the linguistic ontologies at the other hand. In the MikroKosmos initiative, an interface ontology is developed for machine translation purposes in the domain of commercial merges and acquisitions of companies [33]. Hence, it is more close to a given conceptual domain, although general concepts are included as well as unrestricted texts are envisaged to be processed. The KOMET project resulted in the "Generalised Upper Model 2.0", where a closer contact with linguistic realisations is maintained: *if there is no specifiable lexicogrammatical consequences for a 'concept', than it does not belong in the Generalised Upper Model* [34 : p5].

Conclusion

Linguistic semantics is studying how literal meaning is grammaticalised. For our purposes in the GALEN project, we must take the opposite view: what literal meaning can be derived from the lexicogrammatical configurations of the languages that are used for term formation in the domain of surgical procedures. As the GALEN ontology is primarily conceptually oriented, an interface ontology has to be developed that remains close enough to the realisations in these languages. The exact distance of this interface ontology with respect to both sides, has still to be investigated. The question whether such an ontology can assist in finding additional conceptual or linguistic knowledge that is not a priori available in the knowledge bases of the system itself, is another interesting question. What is not a question, but rather a fact, is that answering these questions would be a tremendous step forward to the development of a huge model of medicine as is envisaged in GALEN.

References

1. Rector AL, Nowlan WA, Glowinski A. Goals for Concept Representation in the GALEN project. In Safran C. (ed). *SCAMC 93 Proceedings*. New York: McGraw-Hill 1993, 414-418.

2. 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.

3. Rector AL, Nowlan WA, Kay S. Conceptual Knowledge: the core of medical information systems. In Lun KC, Degoulet P, Piemme TE, Rienhoff O (eds.). *MEDINFO 92 Proceedings*. Amsterdam: North - Holland 1992, 1420-1426.

4. Rector AL. Compositional models of medical concepts: towards re-usable application independent medical terminologies. In Barahona P & Christensen JP (eds.) *Knowledge and decisions in health telematics*. Amsterdam: IOS Press 1994, 133-142.

5. Ceusters W, Deville G, Buekens F. The chimera of purpose- and language-independent concept systems in healthcare. In Barahona P, Veloso M, Bryant J (eds.) *MIE 94 Proceedings* 1994, 208-212.

6. Ceusters W, Deville G, De Moor G. Automated extraction of neurosurgical procedure expressions from full text reports: the Multi-TALE experience. In Brender J, Christensen JP, Scherrer J-R, McNair P (eds.) *MIE 96 Proceedings*. Amsterdam: IOS Press 1996, 154-158.

7. Ceusters W, Deville G. A mixed syntactic-semantic grammar for the analysis of neurosurgical procedure reports: the Multi-TALE experience. In Sevens C, De Moor G (eds.) *MIC'96 Proceedings*, 1996, 59-68.

8. Ceusters W, Lovis C, Rector A, Baud R. Natural language processing tools for the computerised patient record: present and future. In P. Waegemann (ed.) *Toward an Electronic Health Record Europe '96 Proceedings*, 1996:294-300.

9. GALEN Consortium. *Guidelines and Recipes for Completing templates*. Internal document VUM02/96 version 1.0.

10. GALEN Consortium. Links and Templates Summary. Internal document VUM/03/96 version 1.0.

11. Ceusters W, Spyns P. From natural language to formal language: when Multi-TALE meets GALEN. (Submitted for MIE 97).

12. Allen J. Natural Language Understanding. Menlo Park: The Benjamin/Cummings Publishing Company Inc 1987.

13. Quine W. Two Dogma's of Empiricism. In Quine W (ed.) *From a logical point of view*, New York, 1953.

14. Searle JR, Kiefer F, Bierwisch M (eds.) *Speech Act Theory and Pragmatics*. Dordrecht: Reidel, 1980.

15. Bach E. *Informal lectures on formal semantics*. Albany, NY: Suny Press, 1989.

16. Rossi-Mori A. Towards a new generation of terminologies and coding systems. In Barahona P & Christensen JP (eds.) *Knowledge and decisions in health telematics*. Amsterdam: IOS Press 1994, 208-212.

17. Jackendoff R. Conceptual semantics. In Eco U et al. (eds.) *Meaning and mental representation*. Bloomington: Indiana University Press 1988, 81-97.

18. Lakoff G. Cogintive semantics. In Eco U et al. (eds.) *Meaning and mental representation*. Bloomington: Indiana University Press 1988, 119-154.

19. Frawley W. Linguistic Semantics. Hilsdale, Hove and London: Lawrence Erlbaum Associates, 1992.

20. de Saussure F. *Course in General Linguistics*. New York: Philisophical Library, 1959.

21. Ogden CK & Richards IA. *The meaning of meaning*. New York: Harcourt, Brace & Co, 1923.

22. Salton G. Automatic thesaurus construction for information retrieval. Information Processing, 71:115-123, North Holland Publishing Co., Amsterdam 1972.

23. Chen H, Lynch KJ, Basu K, Ng T. Generating, integrating, and activating thesauri for concept-based document retrieval. *IEEE EXPERT, Special Series on Artificial Intelligence in Text-Based Information Systems*, 1993: 8 (2):25-34.

24. Ingenerf J. Taxonomic vocabularies in medicine: the intention of usage determines different established structures. In Greenes RA, Peterson HE, Protty DJ (eds.). *MEDINFO 95 Proceedings*. Amsterdam: North - Holland 1995, 136-139.

25. Bernauer J, Franz M, Schoop M, Schoop D, Pretschner DP. The compositional approach for representing medical concept systems. In Greenes RA, Peterson HE, Protty DJ (eds.). *MEDINFO 95 Proceedings*. Amsterdam: North - Holland 1995, 70-74.

26. Oehrle RT, Bach E, Wheeler D (eds.) *Categorial Grammars and Natural Language Structures*. Dordrecht: Reidel, 1988.

27. 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.

28. Bateman JA. Ontology construction and natural language. *In Proc. International Workshop on Formal Ontology*. Padua, Italy, 1993, 83-93.

29. Kripke S. Naming and Necessity. In Davidson D & Harman G (eds.) *Semantics of natural language*. Dordrecht: Reidel, 1972, 253-355.

30. Welsh C. *On the non-existence of natural kind terms as a linguistically relevant category*. Paper presented at the Liguistic Society of America, New Orleans, LA, 1988.

31. 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.

32 . Deville G, Ceusters W. A multi-dimensional view on natural language modelling in medicine: identifying key-features for successful applications. Supplementary paper in *Proceedings of the Third International Working Conference of IMIA WG6*, Geneva, 1994.

33. Mahesh K. *Ontology Development for Machine Translation: ideology and methodology*. Technical Report MCCS-96-292, Computing Research Laboratory, New Mexico State University, Las Cruces, NM, 1996.

34. Bateman JA, Henschel R, Rinaldi F. *Generalized upper model 2.0*. Technical report, GMD/Institute für Integrierte Publikations- und Informationssysteme, Darmstadt, Germany, 1995.