When definitions are not enough

Michal Boleslav Měchura

Fiontar, Dublin City University

keywords: multi-word terms, compound terms, dependency

1. Introduction

Bilingual terminologists, whose job it is to create target-language equivalents for terms supplied to them in a source language – such as deciding what *wage adjustment* should be called in German, Welsh or Maltese – sometimes find themselves in a situation when the intended meaning of the term is not known to them in its entirety, resulting in misleading coinages. For example, the term *natural language processing* was once rendered in Irish in the Irish National Terminology Database as *próiseáil i dteanga nádúrtha*. This literally means 'processing in a natural language' and it is a misleading attempt at a translation because it implies that natural language is the medium in which some processing occurs. This is not the intended meaning of *natural language processing*. The term actually denotes a situation where natural language is the object of processing – it is the thing being processed. Thus, the appropriate Irish translation should be *próiseáil teanga nádúrtha*, literally 'processing of natural language'. The misunderstanding was eventually noticed by a subject-area expert and corrected, but the problem is a larger one. It concerns situations when a multi-word term is to be rendered in a target language, the translations for the individual words are available but they are combined incorrectly, in contrary to the intended meaning.

Another example is *rich site summary*, a term used in the context of websites and internet applications and better known under its abbreviation *RSS*.[1] The terminologist tasked with the coining an equivalent for this term in a target language may already know how to translate the individual words – for example, she may already know that the sense of *site* invoked here is that of a website rather than a building site, and that *rich* means 'complexly structured' rather than 'possessing a lot of money'. But knowing this is not enough. If the terminologist is not a

[1] The term is actually what's popularly called a *backronym*, a common phenomenon in IT terminology when the abbreviation appears first and the full form is invented later. The form *rich site summary* is one of several often quoted "origins" of the abbreviation (*really simple syndication* is another), but such issues are irrelevant to the discussion here.

subject-area expert, she will probably not know how the words are to be combined: is it ‘a rich summary of a site’ or ‘a summary of a rich site’? Or is it perhaps both? This knowledge is not encoded in the term explicitly.

2. Transitivity, modification and evocation

The problem illustrated here is caused by the terminologist’s lack of knowledge about what I will call transitivity, modification and evocation.

- The concept of **transitivity** refers to the roles that participants play with respect to a process. Roughly speaking, transitivity answers questions of “who does what to whom”. An example of a statement about transitivity is “something processes language” which is a true statement about the term *natural language processing* – in contrast to another possible statement, “something processes something in a language”, which is not a true statement about that term.
- The concept of **modification** refers to the relationship between participants and their modifiers, which are typically adjectives or nouns. An example of a statement about modification is “*natural* modifies *language*”. This is a true statement about the term *natural language processing* – unlike the statement “*natural* modifies *processing*”, which is false.
- The concept of **evocation** answers the question “which sense is evoked by this word?” Words often have several senses but only one of them is evoked in a given multi-word term. In the term *rich site summary*, the sense of *site* is ‘website’ rather than ‘building site’.

For any multi-word term, a number of statements about transitivity, modification and evocation can be made and ideally, they should be known to the terminologist if she is to produce an adequate rendition in the target language. Note that languages differ in how explicitly they encode facts of transitivity and modification. Crucially English, the source language for much of terminology work, leaves many of those facts unexpressed. In *rich site summary*, there is no explicit indication whether *rich* modifies *site* or *summary*. Both interpretations are syntactically possible, the term is ambiguous and one simply needs to be a subject-area expert or have access to one to know which is the intended meaning. Languages other than English often do encode such facts explicitly by means of inflection, agreement or prepositions, forcing the terminologist to decide which interpretation is the intended one in order to be able to coin a rendition in the target language.

3. When a definition is not enough

All this is certainly not a revelation to seasoned terminologists. It is common practice in terminology work to solicit the input of subject-area experts, precisely to avoid the kind of problems illustrated here. The expert's knowledge is usually made available in the form of a definition, and definitions do indeed often clarify ambiguities as to transitivity, modification and evocation. A good definition of *natural language processing* will typically contain enough information for the terminologist to infer that language is the thing being processed rather than the medium in which the processing happens.

Still, even perfectly good definitions sometimes leave a lot of these questions unanswered. A typical definition of *rich site summary* will say something to the effect that it is a file format for making content from one website available to another website or to a computer program. That is a good definition in the sense that it accurately explains what the concept denotes, but unfortunately it mentions neither *rich* nor *summary* nor any of their synonyms, making it impossible to infer facts about the English term's transitivity, modification and evocation – leaving the terminologist in the curious situation of knowing what the term denotes but still unable to proffer an adequate rendition in the target language.

In theory, the terminologist may decide to ignore the source-language term altogether and coin a term based purely on her understanding of what the definition denotes. Such an approach would result in a situation when the terminologist decides that 'content sharing format' or some such rendition will be the target-language term: a wording which takes a completely different route to express the same idea. While this is certainly possible, it is not the approach usually taken. In most cases when target-language equivalents are to be coined for multi-word terms, terminologists prefer for the target-language term to **echo** the internal make-up of the source-language term more or less closely – so that if the source term includes the words *rich* and *site* and *summary*, then the target term should include their appropriate translation equivalents and combine them to the same effect as the source term does. Whether this insistence on fidelity to the source language is a good practice or not is a matter of opinion, but its presence in the routine practice of terminology is a matter of fact. Crucially, in order to achieve the echoing effect, knowledge of the source term's transitivity, modification and evocation is needed.

In other words, definitions are helpful in explaining what a term means but, in the case of multi-word terms, they do not explicate how that meaning can actually be derived compositionally from the individual words. To understand that, facts about the term's transitivity, modification and evocation need to be made explicit.

Strangely, paying attention to the internal structure of multi-word terms does not seem to appear on the agenda of terminology. When the relationship between term and concept is investigated in works on the theory of terminology, it is not uncommon to advocate "the complete dependence on definitions as the only access point and bridge between concept and term" (Sager 1998:261). This paper argues that this is unreasonable and that alongside definitions, a term's internal structure must be an additional "access point and bridge". A compositional analysis of the formal structure of a term can sometimes reveal useful insights into the meaning of the term. While definitions do rightfully occupy a privileged position in terminology in the sense that they can adjust and even override the conclusions from such compositional analysis, they do nonetheless leave certain questions unanswered (as demonstrated by *rich site summary*) that compositional analysis can compensate for.

4. Introducing compositional term diagrams

Hopefully, the previous sections have demonstrated that multi-word terms have an internal structure which is not always apparent but which, nonetheless, is of considerable interest to terminologists. Any instance of language such as a sentence, a phrase or a term appears on the surface as merely a linear sequence of tokens (words). The problem is that these sequences have an internal structure which is not explicitly encoded in them and which needs to be inferred. In linguistics, the process of inferring structure from a linear sequence of tokens is called parsing and the result is usually a syntax tree.

For the purposes of terminology, I propose a formalism called **compositional term diagram** (CTD). CTD is a notation for analysing multi-word terms in order to discover facts about their transitivity, modification and evocation. A CTD is a multi-levelled list of statements which, similarly to a syntax tree, explicates the internal structure of the term. Several examples CTDs follow.

(1) *natural language processing*

```
processing (noun)
  - Q: WHAT IS BEING  PROCESSED?
    A: language (noun)
        - Q: WHAT KIND OF LANGUAGE?
          A: natural (adjective, ‘produced naturally, not artificial’)
```

(2) *rich site summary*

```
summary (noun)
  - Q: WHAT IS BEING SUMMARIZED?
    A: site (noun, ‘website’)
        - Q: WHAT KIND OF SITE?
        - A: rich (adjective, ‘complexly structured’)
```

(3) *database efficiency assessment method*

```
method (noun)
  - Q: A METHOD FOR DOING WHAT?
    A: assessment (noun)
        - Q: WHAT IS BEING ASSESSED?
          A: efficiency (noun)
              - Q: THE EFFICIENCY OF WHAT?
                A: databases (noun, plural)
```

(4) *National Asset Management Agency*

```
agency (noun)
  - Q: WHAT KIND OF AGENCY?
    A: national (adjective)
  - Q: WHAT DOES THE AGENCY DO?
    A: management (noun)
        - Q: WHAT IS BEING MANAGED?
          A: assets (noun, plural)
```

CTDs have the following properties:

- A CTD consists of nodes (underlined). The words correspond to words in the term (and optionally also to other items, such as prefixes). A CTD is hierarchical, like a syntax tree: some nodes are dependants of other nodes. For example, in (1), *natural* is a dependant of *language* and *language* is a dependant of *processing*. A node can have zero, one or more dependants. Like a syntax tree, a CTD has a single node at the top.
- By giving answers to questions such as “what kind of agency?” and “what is being

processed?”, the CTD reveals facts about transitivity and modification in the term.

- The optional paraphrases in brackets reveal facts about evocation in the term: they state explicitly which sense of the word is evoked in the term.

CTDs are structured similarly to syntax trees but are designed to be readily readable by humans, even without training in formal syntax. It is expected that a source-language terminologist, assisted by a domain-area expert, will be able to compose a CTD for each term before passing it on to target-language terminologists (along with a definition, information about the concept’s domain, its relations to other concepts, and any other relevant data). The target-language terminologists will then be able to use the information recorded in the CTD to coin terms in the target languages without running the risk of combining the individual words incorrectly.

If CTDs are used in the practice of multilingual terminology, they will likely prevent miscoinages caused by lack of information about transitivity, modification and evocation. Looking at the CTD in (1), the target-language terminologist will notice that *natural* modifies *language* and not *processing*. She will also know what the nature of the relationship between *language* and *processing* is: language is the thing being processed, not a medium in which processing occurs. Thus, the mistake mentioned at the beginning of this paper will be avoided.

Formally, CTDs are a form of syntax trees called **dependency trees**. Dependency trees are one of the two competing diagramming devices used in formal syntax studies for analysing the structure of multi-word units such as phrases and clauses (the other device being **phrase-structure trees**). Several flavours of dependency trees have been used in literature for the description of countless languages (Meľčuk 1988 for Russian, French, English and others; Sgall *et al.* 1986 for Czech and English; Hudson 1990 for English). For a broad overview of dependency-based approaches to syntax, see Meľčuk (1988).

The bulleted-list notation used for CTDs in this paper is merely a re-representation of what is internally a dependency tree. For example, the multi-word term in (4) could just as well be represented as a dependency tree *à la* Tesnière (1959) in Figure 1 or as directed arcs *à la* Matthews (1981) in Figure 2.

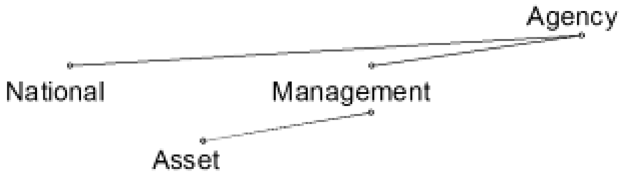

Figure 1: A dependency syntax tree

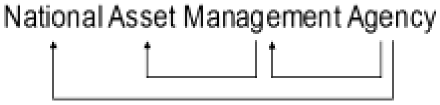

Figure 2: Dependency diagrammed as directed arcs

In dependency syntax, a **dependency** is a directed relation between a **head** and a **dependant**. The relations can be labelled, as indeed they are in CDTs: for example in (4), there is a directed relation between *agency* (the head) and *management* (the dependant) and this relation is labelled with the question "what does the agency do".

5. CTDs and ambiguity

The previous sections have introduced CTDs and explained their formal properties. This section will now demonstrate how CTDs can be can be applied to solve one particular problem in terminology: the problem of internal structural ambiguity in multi-word terms.

Human language is occasionally ambiguous. In linguistics, the effect of attempting to parse an ambiguous instance of language is that more than one syntax tree is produced. In terminology, the internal structure of a multi-word term may also be ambiguous and the term may yield more than one CTD. In some cases, only one of the CTDs is "correct" (corresponds to the intended interpretation) – we can call such cases **false ambiguity**. In other cases, two or more CTDs may be "correct" at the same time – we can call such cases **genuine ambiguity**.

For an example of the former case (false ambiguity) consider the term *descriptive translation studies*. It is possible to construct two CTDs for this term:

(5) *descriptive translation studies*

```
studies (noun, plural)
 - Q: WHAT IS BEING STUDIED?
   A: translation (noun)
 - Q: WHAT KIND OF STUDIES?
   A: descriptive (adjective)
```

(6) *descriptive translation studies*

```
studies (noun, plural)
 - Q: WHAT IS BEING STUDIED?
   A: translation (noun)
       - Q: WHAT KIND OF TRANSLATION?
         A: descriptive (adjective)
```

The CTD in (5) can be paraphrased as ‘descriptive studies of translation’ and the CTD in (6) as ‘studies of descriptive translation’. Upon consultation with a domain-area expert, it becomes obvious that the former interpretation is the intended one while the latter is nonsensical (there is no such thing as “descriptive translation”). This case is similar to the *natural language processing* and *rich site summary* cases: several CTDs are possible but only one is “correct”.

For an example of genuine ambiguity, consider the term *conditional jump instruction*. This is a term from computer programming and denotes a type of instruction given to a computer as part of a computer program. The computer is supposed to “jump” to another place in the program if certain conditions are met, otherwise continue without jumping. Two different CTDs can be constructed:

(7) *conditional jump instruction*

```
instruction (noun)
 - Q: AN INSTRUCTION TO DO WHAT?
   A: jump (noun)
 - Q: WHAT KIND OF INSTRUCTION?
   A: conditional (adjective)
```

(8) *conditional jump instruction*

```
instruction (noun)
 - Q: AN INSTRUCTION TO DO WHAT?
   A: jump (noun)
       - Q: WHAT KIND OF JUMP?
         A: conditional (adjective)
```

The CTD in (7) can be paraphrased as ‘a conditional instruction to do a jump’ and the CTD in (8) as ‘an instruction to do a conditional jump’. Upon consultation with a domain-area expert, it turns out that both are equally plausible: both interpretations are compatible with what the term means and how it is used in computer programming. In the former interpretation, the instruction is only executed by the computer if certain conditions are met, and it tells the computer to jump. In the second interpretation the instruction is executed always and it tells the computer to jump if certain conditions are met. It is an extremely subtle difference, the effect is the same in both cases: the computer jumps if certain conditions are met. Therefore, both CTDs are "correct" in the sense that they are both compatible with the term's meaning as understood by subject-area experts. The term's internal structure is genuinely ambiguous.

Cases of genuine structural ambiguity are probably very rare. When they do occur, they probably occur more often in poorly inflected languages such as English and less often in richly inflected languages. Therefore, when translating an English term such as *conditional jump instruction* into other languages, the target-language terminologist typically has to commit to one or the other interpretation. Interestingly, different target languages can decide in different ways. In IATE, the European Union's multilingual terminology database, *conditional jump instruction* is translated into French as (9) and into German as (10).

(9) *instruction de saut contionnel*
instruction.*fem* of jump.*masc* conditional.*masc*
'an instruction to do a conditional jump'

(10) *bedingter Sprungbefehl*
conditional jump-command
'a conditional instruction to do a jump'

In the French, the gender agreement between *saut* 'jump' and *conditionnel* 'conditional' indicates that interpretation (7) was chosen. In the German, the compounding of *Sprung* 'jump' and *Befehl* 'command' indicates that interpretation (8) was chosen. This is an example of how the internal structural ambiguity of a source term can introduce cross-linguistic inconsistencies among target terms. But structural ambiguity can introduce inconsistency even with a single target language when different interpretations are chosen for analogous terms. Consider the pair of terms *conditional jump instruction* and *conditional stop instruction*. They are isomorphic and they are both genuinely ambiguous in their internal structure: each yields two possible CTDs and they are both "correct". In the Irish National Terminology Database, a different interpretation has been chosen in each case:

(11) *treoir léime coinníollaí*
instruction.*nominative* jump.*genitive* conditional.*genitive*
'an instruction to do a conditional jump'

(12) *stopthreoir choinníollach*
stop-instruction conditional
'a conditional instruction to do a stop'

In (11), the agreement in case between *léime* 'of jump' and *coinníollaí* 'of conditional' indicates that *conditional* modifies *jump*. In (12), the compounding of *stop* 'stop' and *treoir* 'instruction' into a single word indicates that *stop* modifies instruction. Different strategies have been chosen for each term as a result of the source terms' internal structural ambiguity.

Whether such an inconsistency is tolerable or whether the same interpretation should have been chosen is debatable and not the subject of this paper. The point is that CTDs give us the tools to analyse such cases in a formal, rigorous way.

6. Conclusion

There is a tendency in terminology work to devote large amounts of attention to the structure of semantic domains: to ontologies, to concept relations ("is a", "is part of") and generally to the structural aspects of units larger than individual terms. While these aspects of terminology work are interesting and important, they treat terms unfairly as unanalysed units, as little pieces of text whose internal structure does not need to be investigated. I hope I have demonstrated in this paper that the **internal** structure of terms is also relevant for terminology work and that it does rightfully fall within the scope of a terminologist's attention. The formalism of compositional term diagrams (CTDs), as proposed here, gives terminologists just the tools to start looking inside terms with a renewed earnestness. When used to analyse the structure of multi-word terms, CTDs have the potential to help terminologists solve real problems.

mechrm@dcu.ie

References

Hudson, R. A. (1991) *English Word Grammar* Oxford: Blackwell

Matthews, P. H. (1981) *Syntax* Cambridge: Cambridge University Press

Meľčuk, I. A. (1988) *Dependency Syntax: Theory and Practice* Albany, NY: SUNY Press

Sager, J. C. (1998) 'Terminology theory' in Baker, M. (ed.) *Routledge Encyclopedia of Translation Studies* London: Routledge, pp. 258–262

Sgall, P.; Hajičová, E; Panevová, J.; Mey, J (1986) *The meaning of the sentence in its semantic and pragmatic aspects* Academia: Prague

Tesnière, L. (1959) *Élements de syntaxe structurale* Paris: Klincksieck. German translation: (1980) *Grundzüge der strukturalen Syntax* Stuttgart: Klett-Cotta

Terminology databases

Irish National Terminology Database: *www.focal.ie*

IATE: *iate.europa.eu*