The Verb in the Terminological Collocations Contribution to the Development of a Morphological Analyser MorphoComp

Rute Costa, Raquel Silva
Universidade Nova de Lisboa - CLUNL
Unit Research – Lexicology, Lexicography and Terminology
Avenida de Berna 26 – C
1069 – 061 Lisboa – Portugal
m.rutecosta@mail.telepac.pt; raq.silva@fcsh.unl.pt

Abstract

Considering that we are observing and describing the behaviour of the terminological units and the terminological collocations, we intend to talk about the value of the verb as a nuclear element of the terminological collocation in the Portuguese language.
So we will empathize the theoretical distinction between multilexemic terminological unit and terminological collocation and the importance of the verbs in corpus for special purposes.
The characteristics of the verbal paradigms are particularly complex. Identifying, describing and formalizing the complexity of the terminological verbs, within linguistics structures in verbal terminological collocation, is an important issue in the construction of a morphological analyser which purpose is to enrich the performance of ExtracTerm, an extractor for Portuguese language. The growing quality and the rigour of the tagging allows the improvement of linguistic rules in order to obtain better results in the automatic extraction of terminological units.
The morphosyntactical, semantic and conceptual analysis of the terminological verbs will increase the capacity of the automatic analysis.

Terminological Collocations

In the framework of complex terminological collocations, many authors have studied the concepts of collocation and phraseology. From a theoretical point of view, we sustain the need to distinguish collocations from phraseologies, differentiating these ones from the multilexemic terminological units, which are structurally complex, and which function is to denominate concepts. In practice, what puzzles the difficulty to distinguish between these designations is the difficulty to identify the linguistic units to which each one of them refers to, since the frontier between cohesive units, semi-cohesive and free units is not always clear.

Terminologists consider, at first sight, that the differences between terminological units and collocations are of a conceptual domain. That's why the first ones are undoubtedly denominatives; while the seconds are constituted by a set of elements, in which one of them exercises a morphosyntactical and/or semantic power of attraction over the other constituents, which in total compose the collocation.

In the context of language for special purposes (LSP), the terminological unit is recognised because the concept to which it refers to is identified. Thus, in its essence the denominations are terms that designate specific concepts from one or various spheres of knowledge, and theoretically are common to the individuals that constitute a specialized communication community.

In the case of special purposes foreign language learning, the specialist has *a priori* the inherent capacity to construct the conceptual systems of his speciality. It is than necessary to acquire knowledge in terms of speciality speeches construction in the foreign language, as well as to acquire the morphosyntactical constructions intrinsic to the referred speeches.

The learning of a speciality language goes through the command of two indissoluble types of knowledge: the conceptual and the linguistic one. The relation between them is fundamental for the construction of speeches about different areas of knowledge.

In the case of terminological collocations, the specialist identifies the concept to which one of the lexemes refers to. The lexeme has the statute of terminological unit and in a certain syntagmatic context attracts one other lexeme that may be terminological. Therefore, the total morpho-syntactical construction is a non-term, considering that its whole generally does not refer

to a concept. In the situations in which the dependency relation between the elements that constitute the construction - which started by being a terminological collocation - solidifies, it occurs a statute change: the terminological collocation looses its statute to acquire the statue of terminological unit. Some terms such as *tirar sangue*, *medir a temperatura* or *medir a febre*, may be taken as examples.

As a result, the terminological collocations can obey to two types of entities: in the first one, the terminological collocation is constituted by two lexemes, in which one of them has the statute of mono or multilexemic term, and the other the statute of non term. In the second entity, the second lexeme may himself also be a mono or multilexemic term. What characterizes these two entities is the fact that the effect of its two combinations results in a non -term.

Verbal Terminological Collocations (CTV)

There are not many works on the verb description in a scientific language context that intend to elaborate tools to construct terminological resources from specialized *corpora*. This is our objective. We intend to focus on the CTV, that occur in specialized *corpora*, and that we have defined as restrictive combinations, constituted by two lexemes that may be interconnected or not, by autonomous gramemes, having one of the lexemes the statute of verb, and being the other a mono or multilexemic term with a denominative function, every time the occur out of his restrictive syntagmatic context.

To identify the CTV we have started from a medicine *corpus* - **MEDICOTEXT**- compiled at the *Linha de Investigação de Lexicologia, Lexicgrarfia e Terminologia of the Centro de Linguística da Universidade Nova de Lisboa*, restricting ourselves to a *sub-corpus* of Imagiology, with 257 666 occurrences. In this sub-*corpus* we have observed the occurrence of two types of verbs: (i) specific medical verbs such as *radiografar, diagnosticar, cateteriza,,…;* (ii) verbs of general vocabulary, as for instance *aumentar, medir, produzir, distinguir,...* These two types of verbs appear in the medical speech, but the second type had a higher frequency. We will consider now, for instance, the verb *diagnosticar*, which frequency is 78 forms non lematized. Among the forms we found, only the infinitive form of the verb, constitutes itself as a basis element, and attracts the selected structures that were select by itself. Its sum forms the terminological collocations. The verb *diagnosticar* belongs to the medical field, and refers clearly to the medical act of recognizing the nature of a disease through diagnosis. Its grammatical structure privileges entities that allow the denomination of symptoms or diseases, which from a terminological point of view are translated in the accomplishment of a terminological unit. We will consider the following examples:

1. [[diagnosticar]*V*[[massas]*N*[pancreáticas]*Adj*]*N*]*CTV*
2.[[diagnosticar]*V*[[carcinomas]*N*[hepatocelulares]*Adj* [pequenos]*Adj*]*N*]*CTV*
3.[[diagnosticar]*V*[o]*Art:def*[[[neurinoma]*N*[intracanalar]*Adj*]*N* [do]*Prep2*[[VIII]*Num*[par]*N*]*N*]*N*] *CTV*

In these examples, the CTV correspond, from a syntactic point of view, to a VS, which V is strictly connected to the selected arguments: *massas pancreáticas*; *carcinomas hepatocelulares pequenos*; o *neurinoma intracanalar do VIII par*. These arguments correspond to NSs that are characterized, conceptually, by the denomination of an entity of medical nature corresponding, the NS to a terminological multilexemic unit. All the NSs subcategorised by the verb *diagnosticar* can not be inventoried. When the information storage is organised in a terminological data base, the terminological verb *diagnosticar* will have the statute of lexicographic entry, exactly at the same level of the denominations that constitute the NSs, previously identified. .

Yet, this reasoning can not be linearly applied to every CTV. Here are some examples:

4. [[aumentar]*V*[a]*Art:def*[dose]*N*] *CTV*
5. [[diminuir]*V*[a]*Art:def*[dose]*N*] *CTV*
6. [[quadruplicar]*V*[a]*Art:def*[dose]*N*] *CTV*

The verbs in these VS structures seem to possess a minor terminological charge, when compared with the verbal term *diagnosticar*, as empirically, we know they are frequent in non-specialized speeches. Nevertheless, after observing the *corpus* we can verify that these verbs seem to activate specific conceptual traces, when selecting a certain type of terminological units. We will see now what happens with the term *dose*.

Dose is defined as a quantity of a medicine that should be taken in one time. It appears to be a subcategorized term, by verbs that permit the attribution of a quantitative value to the concept of *dose*, and therefore interfering in its definition.

But we also found collocations such as: *optimizar a dose*, *variar a dose*, *administrar a dose*. While the verb *administrar* behaves as a generic, which means that the quantity of the *dose*

is not specified, the verbs *optimizar* and *variar* allow the introduction of generic variables in the *dose* to be taken. In the examples 1, 2 and 3 we have terminological collocations in which the verb, as well as its argumental selection, corresponds to multilexemic terminological units, while in the examples 4, 5 and 6 only the argumental selection of the verb corresponds to a monolexemical terminological unit.

MorphoComp: Automatic Treatment of CTV

Research on CTV follows the development of the **ExtracTerm** - multilexemic terminological units extractor - intending to improve the linguistic description of the terminogenic structures, as well as, to increase the performance of the rules, and the upgrading of the extraction quality. The **ExtracTerm** was conceived in the context of the research taken by the *Linha de Investigação de Lexicologia, Lexicogarfia e Terminoloia of the Centro de Linguística of Universidade Nova de Lisboa* (CLUNL), and has at present the following functions: (i) tagging of the corpus, which corresponds to the attribution of metalinguistic tags that a certain form may possess, independently of the context in which it occurs; (ii) application of disambiguization rules. The **ExtracTerm** applies disambiguization rules which purpose is to annul the multitags, in order to proceed to the application of the following rules: (iii) application of the recognising rules - which consists in the identification of predefined structures that may assume the multilexemic terminological units; (iv) extraction of the multilexemic units.

The **ExtracTerm** works from a tagged dictionary that attributes automatically, tags to the *corpus* and allows the multitagging of each form. From a basis typology, as a result of the observation of the *corpus* - [N1+Adj1]*N* ; [N1+N1]*N*; [N1+Np]*N ;* [N1+Prep+N2]*N;* [N1+Prep+Np1]*N;* [N1+Sigla] *N;* [Sigla1+Sigla2]*N* - we have concluded that the terminological units have these structures as a basis and so the multilexemic terminological units are the result of the combinations these stuctures:

7. [N + Adj] ***N***
*[*radiografia ***N*** mamária ***Adj*** *]****N***
8. [[N + Adj] ***N****+ Adj]* ***N***
*[*ressonânica ***N*** magnética ***Adj]*** ***N*** nuclear ***Adj]N***

Prior to the application of the typology, is necessary to collect the ambiguities of the multitaggs. To exemplify, we created rules to the following situations: distinction between the definite article and the possessive adjective (pronome possessivo), and the demonstrative pronoun and or the preposition:

[Def:art][Pron:poss][Dem:pron][Prep1:a]>[N:f:s]@[Def:art]

In order to make the automatic extraction of the previously observed type of construction – the structures V + N, in which the N corresponds to a mono or multilexemic terminological unit – we decided to treat the medical specialized language verbs differently from the multidominium verbs, which more regularly assume the Portuguese normal verbal flexion rules.

We have observed that the medical specialized language verbs are less subdued to the verbal flexion forms. These verbs assume more often the infinitive forms, the nominal forms of the verb (present participle and past participle) and the 3rd persons' singular and plural of the indicative. We will than restrict our description to the methodology used to analyse the CTVs, in which the verb presents conceptual and semantic characteristics that permit its classification as a terminological verb. We consider the examples based on the verbs *diagnosticar* and *cateterizar*:

9.a.[[diagnosticar]*V:inf:1*[[carcinomas]*N* [hepatocelulares]*Adj*]*N*]*CTV*
b.[[cateterizar]*V:inf:1*[vasos]*N*]*CTV*

10.a. [[diagnosticando]*V:Ger:1*[o]*art:def*[shunt]*N*] *CTV*
b.[[cateteriza]*V:cj:p:ind:3p:s*[a]*art:def*[[veia]*N*[supra-hepática]*Adj*[direita]*Adj*]*N*] *CTV*
c.[[diagnosticada]*V:Pp::f:s*[no]*Prep2:no:m:s*[[radiograma]*N*[de]*Prep1*[perfil]*N*]*N*] *CTV*

We have added new rules to the existing ones, for the extraction of multilexemic terminological units. We have taken into account all the structures that are already present in dictionaries of typologies which underlie the extractor, as well as the existing rules of disambiguization. As an exemplification, we present some restriction rules of selection, considering new identified patterns as the ones that follow:

[[V:Inf:1]+[N]] *CTV*
[[V:Ger:1]+[N] *CTV*
[[V:cj]+[N]] *CTV*
[[V:Pp]+[N]] *CTV*

The **MorphoComp** project has the purpose to develop tools in the framework of the computational morphology, applied to the *corpora*

for special purposes. One of the modules present in the **MorphoComp**, is the morphological analyser **Morphos**, conceived to complete the automatic extraction that is being applied to a Imagiology *corpus*. The **MorphoComp** predicts the constitution of a module that goes beyond the morphological information that describes the grammatical properties of the verbs, and has the function of adding syntactical information related to the categorization and sub categorization of those verbs. As so, we foresee the introduction of information of conceptual and/or semantic type, related to the verbs that assume terminological value, such as *diagnosticar* e *cateterizar*.

The terminological description operated in this type of verbs contains three types of information.: (i) intralexical information, which has to do with morphological, syntactical and semantic aspects; (ii) interlexical information, which has to do with information that describes the relations between lexical units, and in this case relations between verbs and nouns; (iii) conceptual information, which permits to establish a direct relation between denominations and objects from the extra linguistic world.

The morphological analyser operates over the *corpus* and works on a tagged verb dictionary basis. The metalinguistic information added, is morfosyntactical and conceptual, and for the time being the latest one is generic. So, the verb *diagnosticar* is identified as follows:

diagnosticar V:inf:1:tr/T

Diagnosticar is a terminological verb of the first conjugation, which is in the infinitive, and is a transitive verb. With this tagging the **Morphos** «knows» that the terminological verb can apply restriction rules previously defined: NP are constituted by multilexemic terminological units that are already mentioned in the **ExtracTerm**.

The formalization ends up in the expression that corresponds to the following learning rule:

diagnosticar V:inf:1:tr/T$N

in which $, the reading signal, indicates the moment when **Morphos** incorporates the Dictionary of Typologies, already operational in **ExtracTerm,** so to immediately proceed to the application of the rule:

diagnosticar V:inf:1:tr/T$N[N + Adj]

extracting a CTV with the following structure: *diagnosticar massas pancreáticas*.

Conclusive Notes

The **ExtracTerm,** as well as the modules of the **MorphoComp** are dynamic tools which permit to alter- modifing and/or adding - morphosintactical conceptual and semantic descriptions. These tools also permit to increase opportunely, the terminogenic structures, and the recognition rules, as well as the learning rules when the systematic observation of special purposes language *corpora* is being done.

The terminological verbs, as well as the CTV in which they occur, are very frequent in the LSP. We believe that a more profound study of their behaviour will permit to optimize the tools that aim to constitute and conceive lexicographical products, with different objectives, as for instance, every kind of terminological resources; tools of automatic translation and for automatic translation; and tools to the learning of automatic language for special purposes, among others.

Bibliography

- Benson, M., Benson E. & Ilson, R (1986), The BBI Combinatory Dictionary of English. A Guide to Word Combinations, Amsterdam / Philadelphia, John Benjamins.
- Costa, Rute (2001), Pressupostos teóricos e metodológicos para a extracção automática de unidades terminológicas multilexémicas. Dissertação de Doutoramento em Terminologia. Universidade Nova de Lisboa.
- L'Homme, M.C. (1998), Définition du statut de verbe en langue de spécialité et sa description lexicographique, In Cahiers de Lexicographie 73(2), pp. 61-84.
- Hausmann, F.J. (1997), Tout est idiomatique dans la langue, In La locution entre langue et usages. Fontenay Saint-Cloud, ENS Éditions, pp.277-290.
- Heid. U. (2001), Collocations in Sublanguage Texts : Extraction from Corpora, In Handbook of Terminology, Volume 2, compiled by Sue Allen Wright, Gerhard Budin, Amsterdam / Philadelphia, John Benjamins, pp. 788 – 808.