Journal of the
International Institute for Terminology Research
– IITF –

TERMINOLOGY SCIENCE & RESEARCH

Vol. 19 (2008)

Publisher:
Address:
Redaktion: International Institute for Terminology Research (IITF)
Secretary General: Christer Laurén

CONTENTS

FOREWORD

Volume 19 (2008) of the Journal Terminology Science and Research contains three of the papers plus invited comments presented at the Terminology Colloquium of the IITF, held on August 28 in connection with the 16th European LSP Symposium in Hamburg under the title New Voices in Terminology and Future Research Directions. The first three papers plus comments have already been published in Volume 18 (2007), and all the papers presented at the workshop plus invited comments will be published in a printed version at the end of 2008 or beginning of 2009.

According to the organisers, the aim of the Hamburg colloquium was to give a new generation of terminology scholars an opportunity to present their research and to reflect on the future directions of the field together with senior colleagues; thus they wanted to continue the tradition of seeking diversified views on terminology and of fostering discussion.

Each paper presented at the workshop was commented on by one or two experienced colleagues. This volume contains two papers by young terminologists, viz. John Jairo Giraldo Ortiz, whose paper is commented by Klaus-Dirks Schmitz, and by Sérgio Barros. Unfortunately, none of the comments on the latter were ready for publication. The third paper is by the experienced terminologist Larissa Alexeeva and is commented by Øivin Andersen and Maria Teresa Musacchio.

We are two editors who cooperate in compiling and preparing the journal: Nina Pilke (University of Vaasa) and Birthe Toft (University of Southern Denmark). Please submit articles to the editorial board via one of our e-mail addresses (see below).

Vasa and Kolding, August 2008

Nina Pilke
nina.pilke@uwasa.fi

Birthe Toft
toft@sitkom.sdu.dk

Larissa Alexeeva

Perm State University, Russia

COGNITIVE VIEW OF METAPHORICAL TERMS

Abstract

The aim of this paper is to consider metaphorical terms from the point of view of cognitive analysis, one of the aspects of modern terminology. The questions addressed in the paper mainly concern the mechanism of metaphor. What I shall argue here is that substitution as mechanism does not reflect the nature of a metaphor. The analysis of metaphorical terms within scientific texts makes us rather to assume a different idea, viz.: there is more evidence to support the view of metaphor as deriving from the prototypical mental model underlying two different concepts, rather than the view of metaphor as the result of blending or substitution mechanisms that underpin transfer accounts of metaphor (source-target domain).

The theoretical foundations for this assumption are the following:

- specificity of terminological metaphor

- semiotic nature of the term as a sign

- differentiation between the concept and the metaphor.

The specificity of terminological metaphor is that it is "a bridge between old and new theories" (MacCormac 1976). In this sense terminological metaphors are knowledge-laden. Scientific knowledge obtains a derived character. It means that most of theories in science have been predicted with a certain degree of probability. Scientific metaphors conveying new knowledge should also be predicted or prototyped (consider: to pump water and to pump electrons). The important question that arises here is as follows: does a scientific metaphor really convey a new knowledge?

From the semiotic point of view the metaphoric term may be regarded as an iconic sign because it denotes the objects having similar properties (Ch. W. Morris Foundations of the Theory of Signs, 1938). For this reason metaphor does not establish a new type of relation, since it is determined by the nature of a semiotic sign.

The most serious question about the cognitive nature of scientific metaphor is connected with the correlation of metaphor and concept. Does the scientific metaphor really express one concept by means of the other? A positive answer will mean the following: a metaphor is not a concept but rather a model of interacting concepts. A negative answer will lead us to the idea that the scientific metaphor is the proper form of verbalization of one prototypical image that reflects the specificity of cognition.

INTRODUCTION

The aim of my paper is to formulate an account of the metaphorical term, which allows an adequate treatment of metaphor in science. I also hope to indicate a cognitive view of the metaphorical term according to which it is possible to suggest a new interpretation of the concept-metaphor problem. By regarding the nature of concept, viewed as a quantum of thought, I try to give an account concerning one of the most crucial issues of metaphorical terms.

The problems of metaphorical terms have been widely discussed in terminology (Boyd, Grinev, Heisenberg, Kuhn, MacCormac, Ortega-y-Gasset). According to Lakoff and Johnson (1980), – metaphor is traditionally regarded as a relationship between a source domain (literal meaning) and a target domain (metaphorical meaning). The central point of Lakoff and Johnson's theory of metaphor is that any concept from the source domain can be used to describe a concept in the target domain. In this sense metaphor is believed to be the result of blending of two concepts.

However, Lakoff's theory of blended concepts in a certain way resembles the mechanism of substitution described in classical theories of metaphor (Black 1980). Terminologists who regard metaphorical terms usually accept a substitution view of metaphor. For me the views of Boyd, MacCormac and Searle, who assume that the mechanism of metaphor is not specifically linguistic, are of great importance.

The theoretical foundation that has guided me towards my investigation is the following:

- science in general, comparing to art and literature, is a self-arranged and self-reflective system, therefore the process of terminologization is its inner inquiry

- scientific metaphors differ from literary metaphors: scientific metaphors are not puzzles, but rather answers to the puzzles, since they are prototypes of new theories, while literary metaphors are typically interaction metaphors

- the main function of a scientific metaphor is "to introduce theoretical terminology where none previously existed" (Boyd 1980: 357)

- metaphorical models belong not to the logic of justification or proof, but to the logic of discovery (Recoeur 1986: 240)

- metaphorical terms characterize properties of concepts which have yet to be discovered.

It is important to realize that some of these theses have been ignored or denied by the researchers of metaphor. As a result, they were not fully discussed and solved.

The questions I will address in my paper are the following:

- differentiation between the concept and the metaphor

- specificity of scientific metaphor

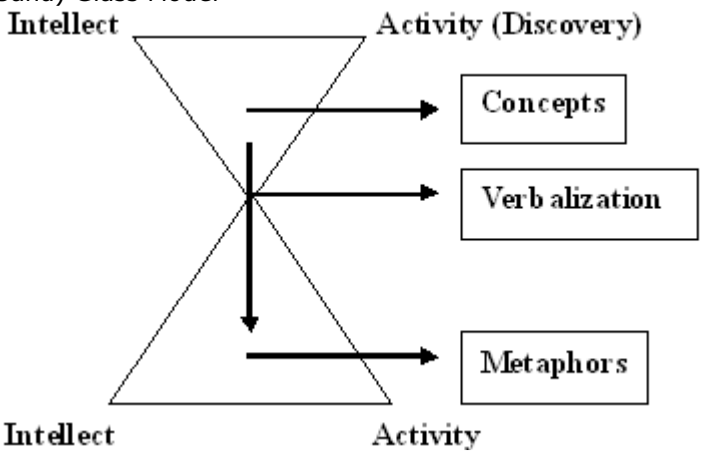
- semiotic nature of the term as a sign.

DIFFERENTIATION BETWEEN THE CONCEPT AND THE METAPHOR

Out of these three questions the most serious one is connected with the correlation of metaphor and concept. The thing is that up to now there is no metaphor theory, based on the type of metaphor – artistic, scientific, ordinary, etc. However, the solution of concept-metaphor problem is directly based on the type of metaphor, since the dependence of metaphor on the type of discourse is very helpful in this case.

In order to discuss this question more thoroughly I have an idea to apply the model of an hour (sand) glass, developing Lakoff's idea that metaphor is not just a matter of language. However, I argue the second part of his statement, connected with the assumption that "human processes are largely metaphorical", and therefore metaphors should be understood as concepts (Lakoff 1980: 6). I would rather affirm that metaphors are **language carriers** of concepts. Consider the model.

The Hour (Sand) Glass Model



I believe that the model of an hourglass will help me in pointing out several features of concept-metaphor relation. The hourglass is taken as a model, since it is capable of affording a starting point for a general account of the nature of the relation between the concept and the metaphor.

On my view, the principle of an hourglass, as a device consisting of two chambers linked by a narrow channel, containing a quantity of sand that takes a special time to trickle to one chamber from the other, is suitable for an interpretation of the problem of concept-metaphor correlation.

In my view, two chambers are associated with two interlinked spheres of man's activity – **intellectual** and **language** spheres. The upper chamber symbolizes mental sphere of man's activity and contains concepts as elements (quantum) of thought. The lower chamber is associated with language activity.

However, two spheres **work together** and are linked by a channel, associated with the process of verbalization. At this point a **concept meets his sign**. In this sense, a metaphor represented by a language form accommodates the concept to be verbalized.

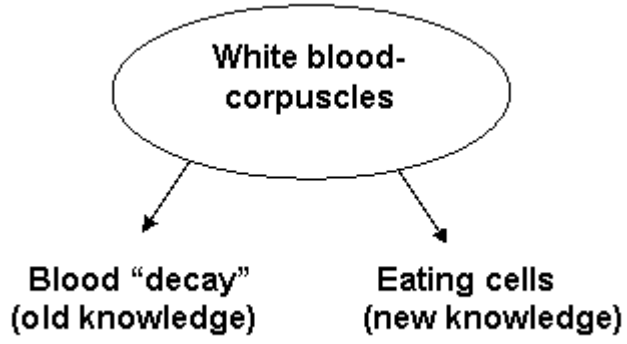
The result of the work of this glass is a metaphorical term. Thus, we assume that the process of terminological metaphorization includes two types of processes: **conceptualization** and **metaphorization**. We grasp the metaphorical meaning of a term only after it is actualized in the language sphere on the basis of its relations with other words.

The most important point in this model is that mental sphere is placed above the language. For me it symbolizes the dominant position of human intellect. It means that in the aspect of the priority, conceptualization is the initial process, and metaphorization comes after it. In this sense, <>metaphorization is derived from conceptualization. The main idea which follows from this model is that up to the moment of metaphorization the discovery has been already done by a researcher.

It means that the concept of a discovered phenomenon has been already formed in the researcher's mind and the only problem for the researcher is how to verbalize it (consider the analogy: a film layer before the developing already contains a print, or a touch, of the reality). In this sense the choice of a metaphor for a newly discovered knowledge is motivated by the already existed knowledge in the form of a concept. We may assume that a researcher carefully selects a proper word (a metaphorical term) in search of an adequate one, which will transfer a new knowledge in the best way. In this sense metaphor (from Gr. meta- "with" and –pherein "to carry") may be regarded as a container of a new scientific knowledge.

I will illustrate my approach to the discussed problem by means of the following example. Ilya Mechnikov (1845-1916), a famous Russian biologist and pathologist in 1882 discovered a unique quality of the human organism – resistance to pathogens. This discovery was later named phagocytosis (Gr. phagein "to eat" kytos "a vessel"). The main idea of Mechnikov's discovery was in the following: the traditional concept of a white blood-corpuscle was that it was a bad symptom of a disease connected with a spoiling of blood.

Ilya Mechnikov proved that white blood-corpuscles engulf bacteria and other harmful particles and in this way help the human organism to subject to phagocytic action. However until 1901 Ilya Mechnikov used in his works a metaphorical term eating cells to describe and to prove his discovery. He modeled his discovery by means of the image of an eating person, who takes some food in, or consumes it.



By this example I shall try to consider metaphor as deriving from the prototypical mental model underlying two different concepts (ordinary and ontological), rather than to view metaphor as the result of blending or substitution mechanisms.

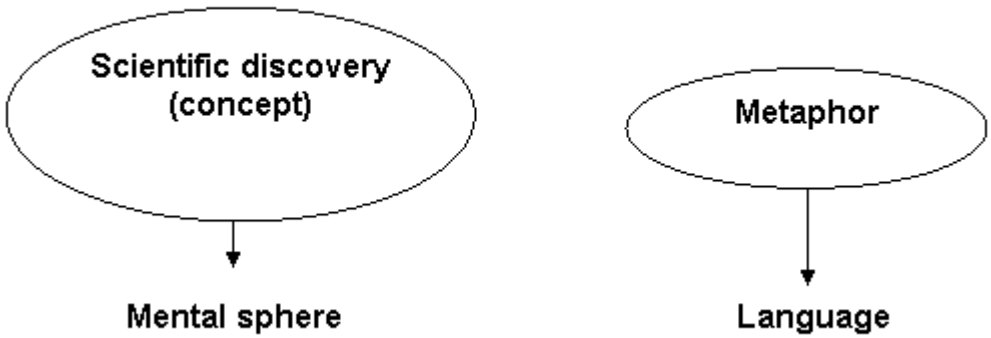
I shall start with the following. In case we have a chance to restore purposes of Mechnikov's choice of a metaphorical term, we should notice that the term eating cells represents several distinctive meanings, corresponding to a new look at the phenomenon of *phagocytosis* as a model of the process, according to which white blood cells ingest pathogen microorganisms.

I believe the best way to explore the concept nature is to observe distinctions of ordinary and specific types of conceptualizations. I want to associate these distinctions with the hourglass model of cognitive explanation of concept-metaphor problem. The distinctive meanings, correlated with different concepts (ordinary and ontological), may be motivated by the following. Consider Table 1:

Table 1

Ordinary observations	Ontological observations
1. Eating as a process is quite discrete: it has the beginning and the end	1. After the disease white blood cells disappear
2. The more food you are offered, the longer is the period of eating	2. The more serious a disease, the greater amount of white blood cells appear
3. All the food is eaten	3. The blood becomes normal

The observation of this table causes the idea that the concepts, structured by the meanings taken from the two columns, are based on the isomorphic meaning: they model the actions of the agents. This model tells exactly what is necessary to comprehend about the new concept. In this case the model does not convey a metaphorical sense. This circumstance helps other scientists to percept a newly born term not as a metaphor, but rather as a natural term (the function of accommodation of a metaphorical term). More than that, in this case the denotate of a new phenomenon exists in reality. Thus, it is quite evident that it is the case of applying one and the same model (prototype) to describe certain phenomena.



Following from this, metaphor (a metaphorical term) serves as a linguistic form for a concept, in other words, it helps a concept to be formed and verbalized. Thus, we may affirm that metaphors are not language dependent, but rather concept dependent phenomena.

In this sense, Langacker's assumption that a semantic structure, functioning as the base for at least one concept, is not correct, taking into consideration the vector of the process of conceptualization in the hourglass model which appears to be quite opposite.

Taking the above said into consideration, it is possible to affirm that a concept is the product of a scientific discovery. Thus, we may conclude that conceptualization and metaphorization as the processes of terminologization are not equal: 1) they are of different nature, 2) metaphor does not create, but rather transfers a new knowledge.

SPECIFICITY OF SCIENTIFIC METAPHOR

New theories of metaphor have moved far beyond the Aristotelian definition of metaphor as the use of **one word** instead of **the other**. "Tenor"-"vehicle" model of metaphor (I.Richards) causes theories of **interaction** of various kinds, regarding the relationship of an image and the abstract meaning.

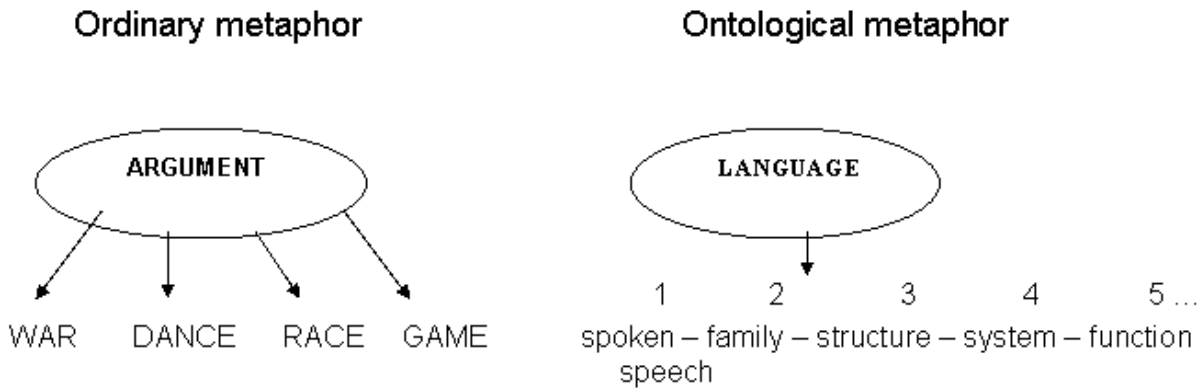
In recent years it was proved that the "essence of metaphor is understanding and experiencing one kind of thing in terms of another" (Lakoff 1980: 5). This may consist of putting two words together, or taking an old word and giving it a new meaning, or creating a new word possessed of some properties which are familiar to us.

In my view, the one who has gone furthest in the direction of studying scientific metaphor is Earl MacCormac, who affirms that the main function of metaphor in science is that of a "root-metaphor" (MacCormac 1976: 93). He suggested that a root metaphor is the most basic assumption about the world or experience that we try to give a description of. More than that, a root metaphor may be applied as a world hypothesis. In his view, imagination is required when one claims to see the world as something other than its obvious character. This imagination act forms the basis of theory construction, for the presumption of a root metaphor is the foundation of a theory (MacCormac 1976: 152). He suggests that scientists who wish to **formulate** new theories that are hypothetical and intelligible have to use metaphors. Simple analogies would be intelligible but not suggestive, while completely new terms would be suggestive but not intelligible. Consider Table 2.

Table 2. The difference between ordinary metaphor and ontological metaphor

Ordinary metaphor	**Ontological metaphor**
All images of metaphor may exist simultaneously	Each image follows the other and provokes a new scientific paradigm
Images of metaphor are formed spontaneously	Images of metaphor are the result of a derived knowledge
Images of metaphor are occasionally created	Images of metaphor are hierarchically dependent, each model causes the shift from one paradigm to another
Images of metaphor are not always connected with the human basic experience	Images of metaphor always are directly connected with the basic human experience

Consider:



Thus, the specificity of the scientific metaphor is that it is "a bridge between old and new theories" (MacCormac 1976: 36). Metaphor makes it possible to develop new meanings in new theories that are intelligible. It is quite obvious that in attempting to describe the unknown, a scientist use terms that are known to us. In this sense terminological metaphors are knowledge-laden. Taking into consideration the difference between the ordinary and the ontological metaphors, it is possible to assume that scientific knowledge obtains a derived character. It means that most of the theories in science have been predicted with a certain degree of probability. Scientific metaphors conveying new knowledge should also be predicted or prototyped (consider: to pump water and to pump electrons).

SEMIOTIC NATURE OF THE TERM AS A SIGN

There is another respect in which metaphorical term may be interpreted not as the result of substitution. From the semiotic point of view the metaphoric term may be regarded as an iconic sign, since it denotes the objects having similar properties (Ch. W. Morris Foundations of the Theory of Signs, 1938). For this reason a metaphor does not establish a new type of relation, since it is determined by the nature of a semiotic sign. More than that, in terms of sign semiosis, terminologization may be viewed as a two staged process: 1) meeting the sign, 2) interpretation of the sign within the frames of other theories. At the first stage the language form joins the already cut concept (nominalization), at the second stage a newly born term starts to be interpreted in other texts (text formation).

CONCLUSION

Within the frames of our paper we tried to find an answer to several important questions: Does the scientific metaphor really convey a new knowledge? Does the scientific metaphor really express one concept by means of the other? A positive answer will mean the following: a metaphor is not a concept but rather a model of interacting concepts. A negative answer will lead us to the idea that the scientific metaphor is the proper form of verbalization of one prototypical image that reflects the specificity of cognition.

The final theses are the following:

1. On the account I am proposing, concepts are intellect-guided phenomena, whereas metaphors are language-actualized units.

2. A cognitive view of scientific metaphors helps to realize the primacy of concepts over metaphors. The hourglass model serves to focus correlation between conceptualization and metaphorization.

3. Purely linguistic approach to metaphors conceals the fact that concepts appear to be self-referential (the term is taken from Searle 2003: 141). It means that concepts are accommodated to language as already built models based on certain prototypes formed by means of distinctive characteristics, obtained in the process of a scientific discovery.

REFERENCES

ALEXEEVA, L.(1998). Term and Metaphor.Perm:Perm State University

BLACK, M. (1980). More about metaphor. Metaphor and Thought.Cambridge, 19-42.

BOYD, R. (1980). Metaphor and Theory Change: What is "Metaphor" a Metaphor for? Metaphor and Thought. Cambridge, 356-408.

LAKOFF, G., JOHNSON, M. (1980).Metaphors We Live by. Chicago, London.

MACCORMAC, Earl R. (1976). Metaphor and Myth in Science and Religion. Duke University Press.

RICOEUR, P.(1986). The Rule of metaphor.Multi-disciplinary studies of the creation of meaning in language.London.

SEARLE, JOHN R. (2003). Consciousness and language. Cambridge University Press.

Maria Teresa Musacchio
Dipartimento di Lingue e Letterature Anglo-Germaniche e Slave
Università di Padova

METAPHOR IN SCIENCE: A REPLY TO LARISSA ALEXEEVA

Abstract

In this reply to "A cognitive view of metaphorical terms" Alexeeva's model and ideas about the use of metaphors in science are compared and contrasted with scientists' views of and comments on their use of metaphors. The objective is to identify possible ways of testing Alexeeva's hypotheses empirically. Based on a short review of some relevant examples in the history of science, metaphors are shown to play an important role in the process of terminologization, while Alexeeva's view that metaphors are just language carriers of concepts is challenged following scientists' reports of how a discovery is made.

INTRODUCTION

The role of metaphor has long been investigated in terminology, particularly with reference to terminologization. Sager (1990: 71-2) includes metaphor among the patterns of term formation, and views the metaphor as a strategy that explores the polysemy of general language since its effect is that something is named after the thing it most resembles. Similarly, Cabré (1998: 100) states that logical relationships between concepts are based on similarity and that one of the semantic methods of term formation consists in modifying the meaning of a term by extending, narrowing or changing its base form (1998: 92-4). Here metaphors are the result of substitution and a process of verbalisation. In the late 1970s, however, a cognitive view of metaphor as interactive tool developed (Lakoff & Johnson 1980) according to which any concept in the source domain can be used in the target domain. Following this line, Temmerman (2000: 70-71), who advocates a socio-cognitive approach to terminology, 'places' metaphor in the mind as she refers to metaphorical reasoning or

> the understanding of a new fact, situation, process or whatever type of category based on the imagined analogy between what one is trying to come to grips with, to understand, and something one knows and understands already. This inventive or creative capacity is made tangible in neo-lexicalisations. These neo-lexicalisations are functional in the process of understanding, which is reflected in language.

In other words, metaphorical reasoning somehow translates into neo-lexicalisations, i.e. results in a process of terminologization.

Based on a range of philosophical and cognitive arguments, Alexeeva subscribes to a cognitive view of metaphorical terms. However, unlike Lakoff and Johnson, and Temmerman, among others, Alexeeva states in her paper that metaphors should not be understood as concepts but rather as language carriers of concepts.

The main points made in the paper in this respect derive from Figures 1 and 2. Figure 1 presents the Hour (Sand) Glass Model used by Alexeeva to differentiate between concept and metaphor; Figure 2 relates scientific discovery (concept) to the mental sphere, metaphor to language and hence regards metaphor as a linguistic actualisation of the concept. In other words, the two chambers are linked by a narrow channel representing the process of verbalisation that translates the concept as the result of the scientist's mental activity into a term which is a consequence of the scientist's language activity.

As metaphors in science relate to the logic of discovery, Figures 1 and 2 appear to follow one from the other. Thus, for Alexeeva, an interactive view of metaphor allows us to construe terminologization as a two-stage process in which the scientist first meets the sign and then interprets it within the framework of other theories.

DISCUSSION

As Andersen states in his comment1, Alexeeva's contribution is very interesting and her exploration of the status and nature of metaphors in science and in scientific terminologization is quite relevant to terminological theory. One problem which Andersen does mention is the lack of empirical testing of Alexeeva's hypotheses, at least so far. In my comments I will focus on this remark as I think that empirical testing raises some challenges for terminology and Alexeeva's views2. In order to do so, I will review the main points which Alexeeva makes, namely the differentiation between the concept and the metaphor, the specificity of scientific metaphor and the semiotic nature of the term as a sign.

There is no denying that philosophers, linguists, psychologists and literary scholars have made fundamental contributions to the understanding of metaphor, but can we take for granted that they understand exactly how metaphor works in science? An empirical testing of Alexeeva's hypotheses should start from what scientists say about how they use metaphors in their own disciplines. Indeed, scientists repeatedly point out that they find it difficult to comment on metaphors in scientific domains other than their own. The reason they give is that there is no scientific method, but as many scientific methods as there are sciences. This is particularly relevant in terminology, as we start from the assumption that terms are coined by domain-experts. I shall return to this point below with reference to terminologization.

The metaphors referred to by Alexeeva seem to be the so-called constitutive metaphors. The distinction between constitutive and exegetical or pedagogical metaphors in science seems to be implied throughout her paper, but is never openly stated. Exegetical or pedagogical metaphors play a role in teaching or explaining theories, while constitutive metaphors are – at least for some time – an irreplaceable part of a scientific theory (Boyd 1979: 359-360). Yet, together with the similar distinction between diaphora and epiphora – whereby scientific metaphor starts off as a diaphora or a kind of suggestion, and ends up becoming an epiphora or description of reality (Vineis 1999) – both play an important role in science. Newton's diaphora of force turned into an epiphora when experiments confirmed his suggestion.

All this also has relevant consequences for terminologization and the diachronic study of metaphorical terms. With reference to the famous exegetical atoms as 'miniature solar systems', Kuhn (1979: 414-415) points out that long after the process of exploring the similarities between the atom and the solar system had ended, Bohr's atom model remained essential to the theory. To provide a further example from physics, the term particle derives from general language where the word designates "a very small quantity of matter" (OED online). This has been used in physics to term metaphorical 'generations' of the smallest components of matter (for a full review, cf. Ahmad 2006). In one of the talks given during his last trip to Italy, Enrico Fermi (1950) pointed out that the term elementary particle had to be understood with reference to state-of-the-art knowledge in the field. Therefore, physicists at the beginning of the 20th century thought the atom was the elementary particle. Then the structure of the atom was discovered and the nucleus was regarded as the elementary particle. In Fermi's times the nucleus was no longer thought of as an elementary particle since its structure had been explored and smaller constituents of matter had been discovered. Fermi's conclusion was that particles are referred to as elementary when their structure is not known; hence the 'generations' identified over the years.

After presenting the hour (sand) glass model (see Figure 1 above) according to which the upper chamber or the sphere of mental activity is linked by a narrow channel to the lower chamber of language activity, Alexeeva states: "The main idea is that up to the moment of metaphorisation the discovery has already been formed in the researcher's mind and the only problem for him is how to verbalize it." Let us look at what empirical evidence we can find in science or what scientists have to say about metaphors as language carriers of concepts. When studying the splitting of atoms, Otto Frisch – before sending the letters to Nature in which he explained the phenomenon he had observed with Lise Meitner – asked a biologist friend what term biologists used to refer to cell splitting. The answer, fission, was thus adopted to designate atom splitting in physics (Maltese 2003: 35). It is interesting to note with reference to Alexeeva's verbalization process that in Italian Fermi first used the somewhat metaphorical term 'scissione dell'atomo' (atom splitting) by analogy with 'scissione delle cellule' (cell splitting) in biology. Only after his move to the US did he use the loan translation 'fissione' in Italian. To take a final example from physics, Fermi – again! – intervened in the discussion to name the small neutral particle now known as neutrino and explained that in Italian one can indicate that something is small or little by adding the suffix –ino or –etto to a noun. As he personally preferred –ino the particle was termed neutrino. The point – both in the case of fission and of neutrino – is that it was not just the 'discovery' that seemed to be in

the scientist's mind, but the metaphor as well, whether this was actualised by language as in the latter example or not as in the former.

According to Alexeeva, conceptualization and metaphorization are not identical processes: 1. they differ in nature, 2. metaphor does not create new knowledge. Montgomery (1996: 135-136) – a geologist writing about science – reports on how Harvey coined the metaphorical term circulation:

> Harvey's choice of "circulation" with regard to the movement of blood through the body, for instance, came directly from an analogy he hit upon to solve the problem of visualising blood flow: 'I began to think', he wrote, using confessional italics, 'whether there might not be a motion, as it were, in a circle'."

Had the discovery already been made? Was it just a problem of coining a term? Perhaps, some ground still had to be covered in order to translate 'a motion in a circle' into 'blood circulation'. Further, Holton (1993: 158) stresses the importance of images in scientists' minds and recalls that Einstein repeatedly stated that he had general images in mind – including metaphors? – and that language played little role in his thinking. Finally, Kuhn (1979: 414) draws a useful distinction between a metaphor and a metaphor-like process and states that scientists may use a metaphor-like process more than metaphor itself – as Einstein seemed to point out in describing his work. With reference to Bohr's metaphor of the atom as a solar system he also mentions that the process of exploring the potential similarities or analogies was not just initial but went 'as far as it could (it has never been completed)' (Kuhn 1979: 415). Could this be taken to mean that a metaphor does provide new knowledge as long as similarities are explored? The following terms in science acquired new meaning over time owing to a metaphor-like process: wave, particle, vein, fault, reaction, immunity, equilibrium, black hole, cold front, dwarf star, dark matter (Montgomery 1996: 135).

Finally, terminologization is viewed by Alexeeva as a two-stage process: 1. meeting the sign, 2. interpretation of the sign within the frames of other theories. From what we have seen above, it is questionable whether scientists meet the sign or what form the sign takes: they may work through a metaphor-like process so that they have an image in mind and try to give a name to it. According to traditional terminological theory, the process is onomasiological, not semasiological, though it is sometimes reversed and proceeds from the term to the concept. However, it seems to me that by rooting metaphor firmly in language, Alexeeva somehow limits its scope in science. Clearly, further examples from many different sciences should be found to test her view. Yet, in socio-cognitive terminology, Temmerman (2000: 71) states that "[metaphorical] neo-lexicalisations are functional in the process of understanding, which is reflected in language", i.e. she seems to place the process in the mind, no matter whether it is also verbalised or not. Moreover, in the interpretation of interactive metaphor by Lakoff and Johnson cognition and language depend on the nature of the pre- and extra-linguistic experience in which cognitive activities take place (Beccaria 2004: 492). Finally, according to Halliday, in science 'these metaphorical expressions are not just another way of saying the same thing. In a certain sense, they present a different view of the world' (1993: 82), i.e. metaphors in science are not just language-actualised concepts, they may imply an altogether different conceptualization.

CONCLUSION

My main purpose here was to outline what could be some of the starting points for an empirical testing of Alexeeva's hypotheses and the fascinating challenges it poses to terminologists. I would like to conclude with Kuhn (1979: 418-419) that perhaps the view of scientific metaphors towards which we should grope in terminology would be one with categories of the mind which could change with time as the accommodation of language (Alexeeva's verbalization of metaphors) and experience (Alexeeva's new concepts as a result of discovery) proceeded. This is a view that would not make metaphor any less real and useful in science and its terminology.

NOTES

[1]I am very grateful to Professor Oivin Andersen for letting me read the draft of his insightful comments on Alexeeva's paper soon after the colloquium in Hamburg.

[2]I thank my colleagues Lorenza Rega, Federica Scarpa and Marella Magris in our research group on terminology at the Scuola Superiore di Lingue Moderne per Interpreti e Traduttori of the University of Trieste for their very useful suggestions for my reply at the colloquium.

REFERENCES

AHMAD K. (2006). Metaphors in the Languages of Science? In Gotti M. and Bhatia V. (eds.) New Trends in Specialized Discourse Analysis. Bern: Peter Lang, 197-220.

ALEXEEVA L.M. (2006). Interaction between terminology and philosophy. In Budin G. et al. (eds.) The Theoretical Foundations of Terminology Comparison between Eastern Europe and Western Countries. Würzburg: Ergon Verlag, 9-19.

BECCARIA G.L. (a cura di) (2004). Dizionario di linguistica. Torino: Einaudi.

BOYD R. (1979). Metaphor and theory change: What is "metaphor" a metaphor for? In Ortony A. (ed.) Metaphor and Thought. Cambridge: Cambridge University Press, pp. 356-408.

CABRÉ M.T. (1998). Terminology. Theory, Methods and Applications. Amsterdam/Philadelphia: John Benjamins.

FERMI E. (1950). Conferenze di fisica atomica: raccolte da professori e assistenti di fisica delle Università di Roma e Milano. Roma: Accademia dei Lincei.

HALLIDAY M.A.K. (1993). Some grammatical problems in scientific English. In Halliday M.A.K. and Martin J.R., Writing Science. Literacy and Discursive Power. London: Falmer, 69-105.

HOLTON G. (1973). L'immaginazione nella scienza. In Le responsabilità della scienza, Bari: Laterza, 145-175.

KUHN T. (1979). Metaphor in science. In Ortony A. (ed.) Metaphor and Thought. Cambridge: Cambridge University Press, 409-419.

LAKOFF G. & JOHNSON M. (1980). Metaphors We Live By. Chicago: University of Chicago Press.

MALTESE G. (2003). Fermi in America. Bologna: Zanichelli.

MONTGOMERY S.L. (1996). The Scientific Voice. Madison: University of Wisconsin Press.

OED (2007). OED Online. http://dictionary.oed.com (site visited in November 2007).

SAGER J.C. (1990). A Practical Course in Terminology Processing. Amsterdam/Philadelphia, John Benjamins.

OED (2007). OED Online. http://dictionary.oed.com (site visited in November 2007).

TEMMERMAN R. (2000). Towards New Ways of Terminology Description. Amsterdam/Philadelphia, John Benjamins.

VINEIS P. (1999). Nel crepuscolo della probabilità. Torino: Einaudi.

Figure 1. The Hour (Sand) Glass Model

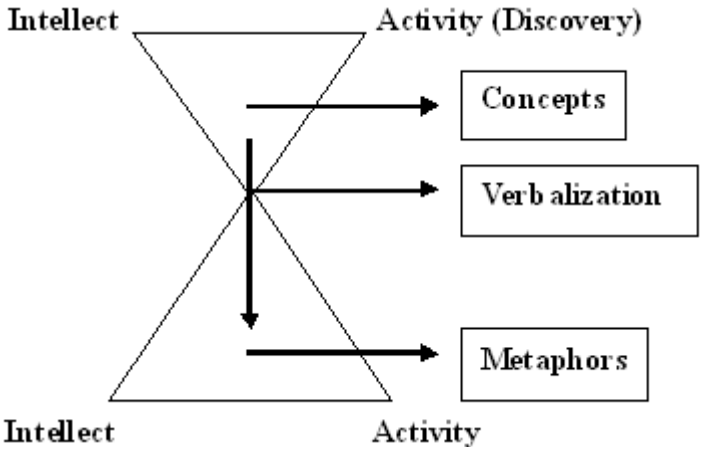


Figure 2. The relation between concept, metaphor and language (the arrow linking concept and metaphor is my addition)

Øivin Andersen
Department of Linguistic, Literary and Aesthetic Studies
University of Bergen
Norway

COMMENTS ON ALEXEEVA'S PAPER "A COGNITIVE VIEW OF METAPHORICAL TERMS"

Abstract

In my comments I will first discuss some of the theoretical concepts which Alexeeva uses, and subsequently I will deal with the empirical and methodological consequences of her theory. Finally, I will reflect on the parallelism between metaphors and models in science.

Alexeeva's contribution is a very interesting and intriguing new voice in modern terminological theory. The status and nature of metaphors have been studied and discussed widely both in cognitive psychology and in linguistics. Alexeeva opposes the traditional terminological view that metaphors are the result of substitution, and advocates an idea that the scientific metaphor is derived from a prototypical mental model underlying two different concepts based on what she calls "co-reference". In order to get this idea across she has constructed an hour glass model.

I do like the idea of making a sharp distinction between the terminological concept and the metaphor. I also like the idea that metaphors are language carriers of concepts. But basically, I think these ideas can be further developed, especially in a terminological context. One of the basic problems is the lack of empirical testing of the hypotheses in question. I also think that several aspects of the theory need to be clarified . In my comments I will first go into some of the theoretical concepts applied by Alexeeva, and subsequently I will deal with the empirical and methodological consequences of her theory. Finally I will reflect on the parallelism between metaphors and models in science.

In several cases, Alexeeva's article applies the concept of prototype without giving references. This theory originated in Cognitive Psychology from the experiments of Eleanor Rosch and her colleagues (Rosch & Lloyd 1978). It soon found its way into linguistic semantics and later also into terminological theory especially via Peter Weissenhofer's dissertation (Weissenhofer 1995). In her theoretical foundations Alexeeva claims that scientific metaphors are "prototypes of new theories". It is not clear what is meant by this as it stands.

According to Rosch & Lloyd (1978), sets of prototypes exist, consisting of members with different statuses. In the normal case, you will have one prototype member (containing all the typical features of the set) and a series of less prototypical and more peripheral members of the same set. In light of the foregoing, how can scientific metaphors be prototypes of new theories? What, in that case, would constitute the non-prototypical members of the set? I think that it is possible to clarify this by using Alexeeva's approach, but it needs to be spelled out and exemplified.

For instance, in her basic hypothesis, Alexeeva supports the view that "the metaphor is derived from the prototypical model underlying two different concepts (based on co-reference)". My questions are: What is the prototypical mental model prototypical of? How does "co-reference" enter the picture here? Towards the end of the article prototypicality is again mentioned: "…the scientific metaphor is the proper form of verbalization of one prototypical image that reflects the specificity of cognition". Again the question is what the prototypical image is prototypical of? These aspects need elaboration, exemplification and clarification, I think.

Discussing the article with Alexeeva at the LSP conference in Hamburg, it was revealed that the terms "prototype" and "prototypicality" were not intended to be used in a theoretical sense in her article. This in itself is quite legitimate, even though these words seem to be central to the basic understanding of her metaphor theory.

Still, I think that it is very important in any scientific setting to distinguish explicitly between the theoretical use of terms and the pretheoretical use of terms. In the former case, a reference to the originator of the theory (in this case Eleanor Rosch or Peter Weissenhofer) should be given; in the latter case it should be explicitly stated that the word is being used in an LGP sense (i.e. pretheoretically). I would recommend Alexeeva to apply the term in a theoretical sense, because it can contribute to clarification of her basic ideas.

The Hour Glass model is interesting, but some theoretical terms remain unclear. The model consists of two chambers where the upper chamber symbolizes the mental sphere of man's activity and contains "concepts as elements (quantum) of thought". It is not clear what is meant by "elements of thought". Are these elements similar to Wüster's characteristic features (Wüster 1985), or are they related to some other conceptual theory? More generally, Alexeeva's contribution does not mention how the new theory is to be integrated into the existing terminological theories, such as the classical Wüsterian theory or the modern cognitive theories of Temmerman and others (Temmerman 2000).

It is obvious that Alexeeva's article is not an empirically oriented contribution, and that theory is focussed on, but I do believe that we can agree on the fact that LSP and terminology is an empirical discipline. This means that it is important to work out the empirical consequences of theoretical claims. As Max Black points out in his book Models and Metaphors from 1962, "even the elementary demand for self-consistency may be violated in subtle ways unless independent tests are available; and what is meant by the reality of the model becomes mysterious" (Black 1962:130).

Karl Popper demanded that theories and hypotheses be empirically tested (Popper 1980). The aim of empirical tests is to obtain relevant data which can lend empirical support to our hypotheses. If these data do not falsify our claims, the hypotheses have been strengthened or corroborated. In order to achieve that, we have to construct a methodology which is derived from our hypotheses. First, we have to ask ourselves whether our theoretical claims are testable in principle, i.e. is it possible at all to imagine an empirical test for the claim in question. If the answer is positive, a fruitful testing method must be constructed.

Let us try to apply this to one of Alexeeva's theoretical claims. Alexeeva claims that establishing a scientific metaphor is a process comprising two consecutive stages: conceptualization is the initial stage, and metaphorization comes after it and is derived from conceptualization.

Is this an empirical claim? If it is, how can it be tested on empirical, terminological data? Alexeeva provides an example from biology and pathology. The example is both interesting and revealing, but it remains an example. It cannot alone serve the function of empirical support for her claim.

Alexeeva says that the essence of metaphor is understanding and experiencing one kind of thing in terms of another. I do agree with this statement. But this is also one of the basic functions of the use of models. In fact the Hour Glass model is at the same time also an elaborated metaphor. Black (1962) has pointed out the strong resemblance between the use of metaphors and the use of models in science, and later, in the 1980s, the American psychologist and logician P.N. Johnson-Laird studied mental models and concluded that they are types of constructions that we create not only in science but also in our daily activities (Johnson-Laird 1987).

These working models (as he calls them) are necessary tools for understanding and coping with the world that surrounds us. Some aspects of the external world are abstract and difficult to understand. In order to grasp them we construct tacit models of them. Sometimes these models turn out to be useless and we readily discard them and construct alternative ones. They clearly have many of the properties of Earl MacCormac's root metaphors (originally developed by Stephen C. Pepper, cf. Pepper 1942). Like Alexeeva's metaphorical terms, mental models can be seen as iconic signs denoting objects having properties similar to the phenomena in the external world that we seek to grasp. In a scientific setting these models are explicated and elaborated, and, as Black pointed out, serve as useful sources of new insights.

To sum up, I do agree with the first two points in Alexeeva's conclusion, i.e. that concepts are intellect-guided phenomena, whereas metaphors, including scientific metaphors, are language-actualized units. But before I am ready to agree on the third point (i.e. that concepts are accommodated to language based on certain prototypes), I should like to see more illustration of the prototype aspects of the theory. I should also like to see how the empirical consequences of her claims can be worked out and integrated

into (or contrasted with) current terminological theories. If this is done successfully, I think that Alexeeva's voice in LSP and terminology would become even stronger than it is today.

REFERENCES

BLACK, M. (1962): Models and Metaphors. Studies in Language and Philosophy. Itacha, NY: Cornell University Press.

JOHNSON-LAIRD, P. N. (1987): Mental Models. Cambridge: Cambridge University Press.

PEPPER, S. C. (1942): World Hypotheses: a Study in Evidence. California: University of California Press.

POPPER, K. R. (1980): The Logic of Scientific Discovery. London: Hutchinson & Co. Publishers Ltd.

ROSCH, E. & LLOYD, B. (1978): Cognition and Categorization. Hillsdale NJ: Lawrence Erlbaum Associates.

TEMMERMAN, R. (2000): Towards New Ways of Terminology Description. The Sociocognitive Approach. Amsterdam/Philadelphia: John Benjamins Publishing Company.

WEISSENHOFER, P. (1995): Conceptology in Terminology Theory, Semantics and Word-formation. Vienna: IITF Series 6. TermNet.

WÜSTER, E. (1985): Einführung in die Allgemeine Terminologielehre und Terminologische Lexikographie. Copenhagen: The LSP Centre, Copenhagen Business School.

Sérgio Barros
Center for Linguistics
New University of Lisbon
Portugal

DEFINING CONTEXTS AND DELIMITING CANDIDATES FOR THE GENERIC RELATION IN A CORPUS OF PORTUGUESE CONSTITUTIONAL LAW"

Abstract

Currently, many researchers in Computational Terminology focus their work on the development of resources capable of structuring terminologies into concept systems. The terminologist is faced with the difficult task of analyzing, filtering and systematizing large quantities of data. Such a need becomes the motivation for a reflection on the interaction of different levels of analysis inherent to terminological units. The aim of the present paper is directed towards the identification of knowledge-rich contexts, where terminological units which are candidates for the generic relation can be delimited. Our approach is based on a corpus of Portuguese Constitutional Law, taking into account defining contexts which included a specific linguistic pattern. The analyzed data form the basis for a typology of contexts for the generic relation in Portuguese Constitutional Law. This research underlines the importance of developing suitable methodologies in Computational Terminology before carrying out natural language processing tasks and prior to developing an application that satisfies the needs of language professionals or domain specialists.

1. INTRODUCTION

Structuring terminologies within companies and governmental institutions is a focus of interest in Terminology, in order to aid language professionals or experts of technical-scientific domains carrying out tasks related with knowledge organization. In this respect, semantic relations established between terminological units can be more explicitly elucidated through concept systems which reflect the way a given technical-scientific domain can be organized. Within the context of language for special purposes, and in connection with knowledge engineering, the applications directed at this type of tasks constitute a valuable resource, optimizing the management of information for specific ends, which could then become faster and more effective.

1.1. SCOPE AND AIM

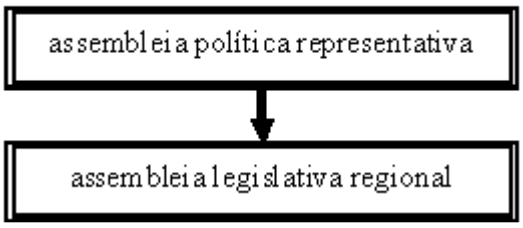
This article corresponds to the initial stage of a research in Terminology regarding the generic relation in the domain of Portuguese Constitutional Law, carried out at the Centro de Linguística1 of Universidade Nova de Lisboa under the supervision of Professor Rute Costa. In the context of specialized knowledge organization, we have defined semantic relations as our object of study and as the starting point for the construction of a concept system in Portuguese Constitutional Law. The immediate objective of this research regards the semi-automatic identification and delimitation of the elements that compose one particular semantic relation: the generic relation. In this article we are mainly concerned with the methodological aspects underlying the automation of such a task, thus stressing relevant theoretical principles, as well as the practical approach we've used.

2. METHODOLOGY

2.1. Theoretical approach

The theoretical component of our research regards the theoretical reflection on our object of study, taking into account the main aspects of the generic relation and its framing in this terminological/terminographical work. More precisely, we were interest in knowing its behavior in a very specific corpus dealing with Law. Within the scope of Terminology work, and in accordance with ISO 1087, a generic relation corresponds to a relation between two concepts, «… where the intension of one of the concepts includes that of the other concept and at least one additional delimiting characteristic…» (1087 ISO 2000:5). Regarding the distinction that Wüster (1998) establishes between logical and ontological relations, the generic relation corresponds to a logical relation, which is based on the degree and type of similarity between two concepts. For example, the sequence «…A revolução é uma força não regulada…»2 presents a relation of subordination between the concept of 'revolução'3 and the concept of 'força não regulada'4, such that the superordinate term is said to have fewer characteristics than the subordinate term. Also, a sequence like «…a assembleia legislativa regional é uma assembleia política representative…»5 shows a hierachical relation between two terms.

Figure 1 –*Hierarquic representation*

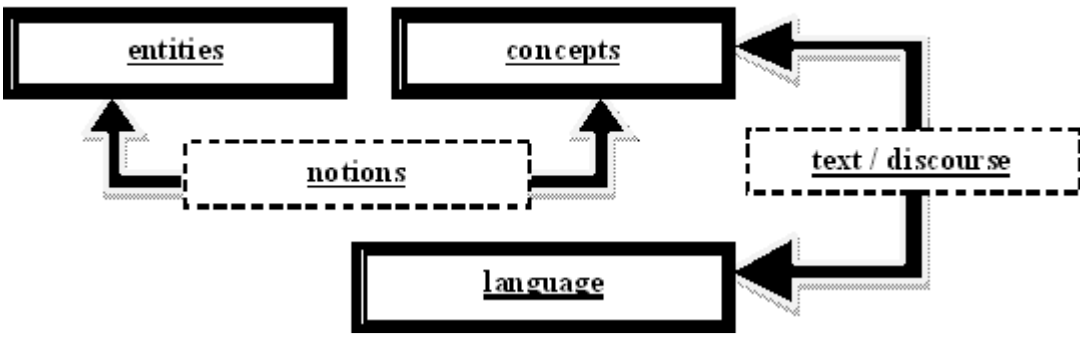


In *Figure 1* we illustrate a possible representation of the hierarchy between the superordinate term 'assembleia política representativa'6 and the subordinate term `assembleia legislativa regional'7, thus assuming there are several representative political assemblies other than the regional legislative assembly. Considering the linguistic units composing this relation, we would like to stress the distinction between term and concept. In fact, in the previous example, there is a logical relation between two terminological units, being that the concept is something which is designated – and not directly accessible – in both terms related via the expression é_uma. This means that 'revolução', like 'força não regulada', does not correspond to a concept but to the linguistic aspect of the terminological unit. This is one of the most relevant aspects of our research, since the defining contexts that we observed in the corpus allow us to delimit term candidates - and not concepts - for the generic relation. In other words, the role that the generic relation plays in our research, at least until this stage, should be understood as a strong indicator of a possible relation between two concepts.

2.1.1. Model of knowledge representation

The theoretical reflection on the terminological practice has a special importance on our research because the terminologist can disperse in the crossing of various levels of analysis subjacent to the terminological unit. Inspired by the model of representation of Monterde Rey (2004) and in the synthesis that the author presents on the authors who are concerned with the way of representing knowledge since Plato and Aristotle until our days, we designed a model of knowledge representation which gives an account of the existing relations between reality, concepts and linguistic units.

Figure 1 – *Knowledge representation model*



Regarding *Figure 2*, we present a synthesis of a model that comprises three main systems and two auxiliary levels. To illustrate the functioning of it, we'll take the example of 'manifestação'[8] throughout the model. This linguistic unit can be seen as a material entity, as a group of people demanding their rights, etc. Therefore, it would be located at the level of a system of entities, where one could place objects of the extra-linguistic reality. Between this level of analysis and the one where we locate the concepts there exists a referential universe where the notion of 'manifestação' is located, being more or less common to a given community. Imagine two people speaking about the Constitution without any of them possessing scientific knowledge in the area of Constitutional Law. Certainly, what they say concerning the concepts of this domain is only a notion of what really is the Constitution. Having two jurists in the same situation would result in a higher level of specialized knowledge, which is less transparent because it makes use of well delimited concepts. The level of analysis of the notions allows us to relate entities with their respective concepts. At the conceptual level, the concept of `manifestation' is created *«… by a unique combination of characteristics…»* (cf. 1087 ISO 2000:2). Concerning immaterial objects, more common within social and human sciences, their characteristics are more difficult to define, more precisely those characteristics that can be observed by simply examining the object and that require no further knowledge about its origin or use (cf. Wüster 1998:46). Then, we have a semiological system where we locate a set of choices, with no syntagmatic organisation, that the individual can make use of for communication: language, linguistic units in its purely lexical facet. At this level we could identify terminological designations such as 'Estado'[9], 'Constituição'[10], 'proporcionalidade'[11], 'igualdade'12 , etc. Between this system and the conceptual system there exists a textual/discursive level which holds the mechanisms that allow the construction of knowledge - the generic relation is one of these mechanisms being visible through recurrent linguistic structures.

2.2. Practical approach

The second component regards acquiring and analysing the initial data that serve as a basis for our study. Following Meyer (2001), we took into account knowledge-rich contexts where it is possible to identify linguistic structures that express semantic relations, thus considering the behavior of the linguistic marker é_uma13. Typically, in Portuguese, this linguistic structure could configure a hierarchical relation between a subordinated term and a superordinate term, as we presented previously in Figure 1. Nevertheless, there are also examples where such linguistic marker establishes a relation between a concept and a delimiting characteristic, as in the example «…o Estado é uma organização diversificada actuante…»14, where 'organização diversificada actuante' is to be considered as a characteristic of 'Estado'.

2.2.1. Material

In order to carry out our research, we constituted a corpus pertaining to the domain of Portuguese Constitutional Law, its dimension in .txt format not exceeding 5Mb, with 31,449 unique words and 777,818 occurrences in total. At an initial stage, we extracted the occurrences of the linguistic marker é_uma, a process which originated 153 defining contexts. Considering sequences as «…O Estado é uma forma aperfeiçoada de ordem…»15 or «…a revolução é uma força não regulada…"16, we verified the cotext adjacent to the linguistic marker identifying the elements that constitute the candidates for the potential generic relation, namely the subordinate and superordinate terms. Systematizing the data at our disposal allows us to elaborate a typology of the generic relation in Portuguese Constitutional Law, thus supporting us in the way of conceiving a method for the semi-automatic identification of the generic relation.

Having observed the defining contexts in which the linguistic to marker é_uma occurs, we observed three main evidences regarding the behavior of the elements that constitute the potential generic relation:

> i) a criterion of contiguity, regarding the linearity and the concatenation of the elements of the generic relation;
> ii) a criterion of order, regarding the positioning of the subordinated and superordinate terms inside the logical structure of this relation;
> iii) and semantic characteristics which contribute more directly or in a more evident manner to the meaning of the linguistic units in relation;

2.2.1.1. Contiguity

The first criterion deals with the concatenation of the linguistic units that are candidates to subordinate and superordinate terms along with the linguistic marker é_uma, like in the example «…O Estado é uma ordem normativa…»17. As we can observe in the example, the generic relation presents a sequential structure among its elements, which are immediately adjacent to the linguistic marker. We designate these sequences as contiguous, in contrast with the non-contiguous sequences, whose subordinate candidate term is not adjacent with é_uma: «… O Estado é um conceito; existe, porque pensado por governantes e governados; e é uma instituição que incorpora uma ideia de Direito…»18. This example illustrates the behavior of the subordinate candidate term in this type of sequences, where a portion of text is intercalated between 'Estado' and the superordinate term 'instituição'.

2.2.1.2. Order

This criterion regards the positioning of the elements that constitute the generic relation within its logical structure. Typically, the subordinate candidate term occurs before the superordinate term, being this its canonical structure: «… A manifestação é uma reunião qualificada…»19. Nevertheless, we observed examples where this order is inverted: «… É uma constante do Direito Constitucional português a unidade do poder político …»20. As we can observe in this example, the subordinate candidate term 'unidade do poder político' occurs after the superordinate term 'constante do Direito Constitutional português' within the logical structure.

2.2.1.3. Semantic characteristics

Within the morphosyntactic structure of the generic relation there are some semantic characteristics, of which two present a bigger relevance: exclusion and anaphora mechanisms.

2.2.1.3.1. Exclusion

While as a generic relation typically is expressed by an assertion between two terminological units via the linguistic marker é_uma, there are sequences where this relation is to be considered as a non-assertion, through the occurrence of an adverb of negation before é_uma: «… O Estado não é uma ordem normativa…»21. The example stresses the necessity of taking into account the sequences where defining expressions like é_uma co-occur with elements other than the terminological units in a generic relation. In comparison with the example of the English linguistic marker is_a, there's the need to filter eventual cases like this, since the negation adverb occurs between "is" and "a", as in x_is_not_a y. In Portuguese, since the negation adverb occurs between the subordinate term and the linguistic marker, the semi-automatic extraction of é_uma includes sequences of exclusion – or non-inclusion – between the two related terms.

2.2.1.3.2. Anaphoric mechanisms

With regard to the anaphoric mechanisms, we can observe several examples where the candidate to subordinate term appears in the form of a pronoun or even without lexical realization. In the first example «… . O primeiro é uma unidade de ordem, …»22, the unit 'primeiro' refers to a term which occurred before in the text, whereas in the second example «… A Constituição do Estado não é processo, mas produto; não é actividade, mas forma de actividade; é uma forma aberta, …»23, we do not even have a linguistic unit immediately before é_uma refering to the subordinate term of 'forma aberta'.

3. METHOD

We present in Table 1 some examples of defining contexts taken from the corpus, where the linguistic marker (LG) is in a pivot position, the subordinated (Sub. Term) and superordinate (Sup. Term) terms to the left and the right of é_uma, respectively, as well as the graphical marks or conjunctions that delimit the beginning (B) and the end (E) of the generic relation.

Table 1 - Examples of candidate sequences to the generic relation taken from the corpus.

ID	B	Sub. Term	LG	Sup. Term	E
1	.	*The socialist democracy*	*is a*	*directed democracy*	,
2	.	*The State*	*is a*	*perfected form of order*	.
3	.	*The revolution*	*is a*	*non-regulated force*	,
4	.	*The Constitution*	*is a*	*frame-order*	,
5	.	*The duty*	*is a*	*passive juridical situation*	,
6	.	*The Constitution*	*is a*	*technique of limitation*	,
7	.	*The objection of conscience*	*is an*	*expression of minority*	;
8	that	*the State*	*is an*	*acting diversified organisation*	,
9	that	*the Constitution*	*is an*	*open order*	,
10	.	*The Portuguese State*	*is a*	*Democratic Republic*	that
11	that	*the magistrate*	*is a*	*spoken law*	or
12	.	*The manifestation*	*is a*	*qualified reunion*	-
13	that	*the English Constitution*	*is an*	*unwriten Constitution*	(

As we can see in Table 1, it is convenient to go over the examples, filtering them and check whether they're valid sequences of a generic relation. Having into consideration the semi-automatic identification of sequences where a generic relation occurs, we are interested in delimiting its constituent elements. In most of the cases, the graphical marks delimit the beginning and/or the end of the generic relation. Let us take example 12) where the subordinated term `manifestation' is correctly delimited by the full stop that occurs before é_uma; also where the superordinate term 'qualified reunion' is delimited by the occurrence of a hyphen after the linguistic marker. A formalization rule like [./que] _ [x] _é_uma_ [y] _ [./,/;/-/que/ou/ (] would account for all the cases shown in Table 1. In terms of future automatic processing, the formalization of these sequences could dispense a previous annotation of the corpus.

4. CONCLUDING REMARKS

In this article we've presented a corpus-based approach for the terminological/terminographical work in order to define the contexts where the generic relation occurs via a specific linguistic marker. Regarding the automation process that we intend to develop, we consider it as an automation axle in which the terminologist gradually reaches higher levels of autonomy in the tasks he develops. At this initial stage,

we proceed with the semi-automatic identification of the elements that, potentially, constitute the generic relation. The corpus we've used is relatively small, functioning as an observatory for our study. On the basis of the systematization of the data, we will be able to consider the hypothesis of implementing a method of automatic identification of the elements that constitute the generic relation, via paralinguistic patterns (cf. Meyer 2001) and/or conjunctions. Thus, at this stage we've achieved a starting point for the construction of a concept system, a sort of taxonomy from where the terminologist can eventually go deeper into the conceptual level of the linguistic units delimited through graphical marks and/or conjunctions. Eventually, a natural development of our research is to enter in the field of the concept modeling.

Generally speaking, we can say that concepts are units of thought or units of knowledge that refer to entities/objects of the extra-linguistic world. Concepts are made of characteristics based upon the properties of the object itself. Such properties are relatively simple to identify when the terminologist is working in a more technical domain. Nevertheless, when it comes to domains belonging to social or human sciences, the distinction between properties and characteristics becomes less evident, since the entities to which concepts refer are essentially immaterial. Regarding this distinction between immaterial and immaterial objects, Nissilä (2006) also points out that «…when it comes to concepts that are based on non-concrete objects, it appears that terminological methods still need to be developed…» (Nissilä 2006:287). A technical-scientific domain more related to social and human sciences, as in the case of Constitutional Law, contrasts with more technical domains regarding the nature of the object of study. As Vignaux (1999) affirms, the text coincides with the observable reality itself: «...L'objet, du moins ce qu'on en croit, fait partie de la réalité, mais il est aussi construit par le langage et produit par l'expérience...» (Vignaux 1999:29).

The model of knowledge representation that we synthesized allows us to conceive a terminological unit under several levels of analysis, whether it's a mere lexical unit, a unit of knowledge, or a referential unit; while reflecting upon the course that the terminologist has to cover between data, information and knowledge in different phases of the research work, depending on the needs. Rute Costa affirms that «…In order to build systems of knowledge representation it is essential to operate simultaneously at the level of the state of things, conceptual networks and semantic structures…» (Costa 2006:81).

4.1. Future work

Considering the material gathered, one should always consider its validation by an expert in a posterior phase so as to verify:

i) cases of metaphor like "… O Homem é uma raça…"24, where 'raça' has the function of a delimiting characteristic rather than of a superordinate term;
ii) the relevance of the comma when delimiting terminological units and defining its extension ("… esta norma de limites materiais é uma norma instrumental, declarativa ou de garantia,…"25), as well as of the conjunction 'ou'26 ("… democracia militante ou protegida é uma contradição in terminis…"27);
iii) anaphora resolution, closely related with the non-contiguous sequences (cf. anaphoric mechanisms in 2.2.1.3.2.);
Meanwhile, the available data allow us to affirm that the work we could develop in the future regards the surveying of the morphosyntactic patterns of the units delimited, bearing in mind the eventual development of machine-learning based methods of automatic identification. On the other hand, it is also important to define the importance of this semantic relation in this domain, before even studying other linguistic structures conveying the generic relation. Other future work regards the creation of a relational database that includes information on the generic relation within Portuguese Constitutional Law, which could also lead to a study of other semantic relations.

NOTES

[1]Center for Linguistics of New University of Lisbon

[2]The revolution is a non-regulated force

[3]revolution

[4]non-regulated force

[5]The regional legislative assembly is a representative political assembly

[6]representative political assembly

[7]regional legislative assembly

[8]Equivalent to the English 'manifestation'

[9]Equivalent to 'State'

[10]Equivalent to 'Constitution'

[11]Equivalent to 'proportionality'

[12]Equivalent to 'equality'

[13]Equivalent to the English marker is_a.

[14]«the State is an acting diversified organisation»

[15]The State is a perfected form of order.

[16]Revolution is a non regulated force.

[17]The State is a normative order.

[18]The State is a concept; it exists because its thought by those who govern and those who are governed; and i tis na institution incorporating an idea of Law.

[19]The manifestation is a qualified reunion.

[20]It is a constant of Portuguese Constitutional Law the unit of political power.

[21]Cf. note 8

[22]The first is a unit of order

[23]The Constitution of the State is not a process but a product; it's not activity but a form of activity; it's an open form

[24]Man is a race

[25]This norm of material limits is an instrumental, declarative or guarantee norm

[26]Equivalent to English 'or'

[27]Protected or militant democracy is a in terminis contradiction

REFERENCES

CABRÉ, T. et al. (2005) "Introduction – Application-driven terminology engineering." Terminology – 11:1, Amsterdam: John Benjamins, 1-19.

COSTA, R. (2006) "Plurality of theoretical approaches to Terminology", in PICHT, E. (ed.) (2006) in Linguistic Insights Vol. 36 - Modern Approaches to terminological theories and applications, Bern: Peter Lang, 77-89.

ISO/FDIS 1087-1 (2000) Terminology work – vocabulary – Part 1: Theory and application, International Standard ISO/FDIS 1087-1:2000.

ISO/FDIS 704 (2000) Terminology work – principles and standards, International Standard ISO/FDIS 704:2000.

MEYER, I. (2001) "Extracting knowledge-rich contexts for terminography" in BOURIGAULT, D. et al. (eds.) Recent Advances in Computational Terminology, Amsterdam/Philadelphia: John Benjamins, 279–302.

NISSILÄ, N. (2006) "Concept Systems in the Balance Sheet", in PICHT, E. (ed.) (2006) in Linguistic Insights Vol. 36, Bern: Peter Lang, 287-300.

NUOPPONEN, A. (1996) "Concept systems and analysis of special language texts", in BUDIN, Gerhard (ed.) Multilingualism in specialist communication, Proceedings of the 10th European LSP-Symposium 1995, Vienna, IITF/TermNet, Vienna, 1069-1078.

NUOPPONEN, A. (2005) "Concept relations – An update of a concept relation classification", in MADSEN et al. (eds.) (2005) TKE 2005 – 7th International Conference on Terminology and Knowledge Engineering, 127-138.

REY, A. M. M. (2004) "Evolución de modelos de formas de representatción del conocimiento a nivel terminológico: propuesta de un modelo actual", in LSP & Professional Communication, Copenhagen: Dansk Selskab for Fagsprog of Fagkommunikation, april 2004. vol.4, nº 1, 49-68.

SAGER, J. (1990) A practical course in Terminology processing, Amsterdam: John Benjamins.

SISSECK, L. (2005) "Terminological knowledge extraction – and machine learning for Danish", in MADSEN et al. (eds.) (2005) Proceedings of the 7th International Conference on Terminology and Knowledge Engineering – TKE 2005, Copenhagen, 385-395.

VIGNAUX, G. (1999) Le demon du classement. Penser et organizer, Paris: Éditions du Seuil.

WÜSTER, E. (1979) "Introducción a la teoría general de la terminología y a la lexicografía terminológica", in CABRÉ, T. (dir.) (1998) Monografies, Barcelona: IULA.

John Jairo Giraldo Ortiz
Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra
Barcelona

DESCRIPTION AND ANALYSIS OF INITIALISMS IN THE GENOMICS AND ENVIRONMENT SUBJECT FIELDS

Abstract

This research paper seeks to achieve two goals: provide linguistic and statistical analyses of the initialisms in the specialized discourse of the domains of Genomics and Environment establish patterns for developing automatic recognition systems to facilitate the building or update of online abbreviation dictionaries. This paper is divided into four main parts: 1) Initialisms - definition and typology; 2) Corpus quantitative analysis; 3) Corpus qualitative analysis, and 4) Initialism recognition patterns.

Key words: Initialism, Terminology, Spanish, Corpus, Genomics, Environment

1. General background

1.1 Terminology and initialisms

We base our research on the Communicative Theory of Terminology (Cabré, 1999), which establishes three different types of specialized knowledge units, i.e., terminological units (or terms), phraseological units, and sentential units.

Cabré maintains that terminological units are classified according to form, function, meaning and origin. With regard to their form, terms can be classified according to the following criteria:

- Number of morphemes. Terms can be simple or complex, e.g.: acid/acidification

- Type of morphemes used in the formation of a complex term. Terms can be derived or compound, e.g.: ulcerous, handball

- Word combination following a specific syntactic structure, e.g.: value-added tax, liquid crystal display, and

- Units of complex origin such as initialisms, which are created by means of shortening processes, e.g.: DNA (Deoxyribonucleic Acid).

According to the abovementioned, an initialism can represent the shortening of a terminological unit. In such a case, the initialism will maintain the same status. That is to say that it will also be considered a terminological unit.

1.2 The effect of initialisms in the specialized discourse: The case of Urology

The study by Bloom (2000: 4) is a good example of the growing use of initialisms in specialized discourse. In the article, Acronyms, abbreviations and initialisms, Bloom presents an interesting diachronic study on initialisms in the most prestigious manuals and journals of Urology. He comes to the

conclusion that initialisms were not so common in these publications at the beginning of the 20th Century. Conversely, they were highly frequent at the end of the century.

Bloom analyzes different journals and manuals published through the 20th Century such as Journal of Urology, British Journal of Urology, The Principles and Practice of Urology, and Urological Surgery. His study shows that from 1900-1950 these publications used only well recognized abbreviations such as cc, mg, or Fig. But in 1950, volume 63 of the Journal of Urology, dedicated almost exclusively to the Transurethral Resection of the Prostate, registered for the first time a timid use of the initialism TUR in a table.

It was not until 1970 that initialisms began to appear continuously in the publications of the field. In fact, that year the Journal of Urology published five articles about Transurethral Resection of the Prostate in which the initialism TUR occurred 34 times. In 1973 the first article of Urology included 21 occurrences of six different initialisms like DNA, RNA and ACTH.

Some years later, in 1990, TUR and TURP were widely used in the publications of the field. Thus, for example, volume 66 of the British Journal of Urology published one article in which TURP appeared 9 times in the text and 6 times in its tables. In addition, 18 initialisms like CVP, RISA, and NS were also used.

In 1996, the British Journal of Urology introduced the key to decipher all the initialisms and abbreviations in its articles. It consisted of a list of the abbreviations and their expansions. It was a measure which the editors implemented since they did not want initialisms and their expansions to co-occur inside the texts. This was a policy that was contrary to the international conventions that pointed out the necessity of the occurrence of the initialism and its expansion at least in the abstract.

Finally, in 2000, the Journal of Urology included a great number of abbreviations. In the abstracts alone there were 47 different abbreviations that occurred 285 times. PSA and BCG were the most frequent ones with 87 and 30 occurrences, respectively.

Despite the generalized use of initialisms, today it is frequent to find letters to editors and articles that criticize openly such an uncontrolled proliferation of initialisms (Morgan, 1985; Green, 1990; Cheng, 1997, 2002a, 2002b, 2005; Walling, 2001; Fallowfield, 2002; Lader, 2002; Fred, 2003; Rowe, 2003; Jack, 2003; Bradley, 2004; Isaacs, 2007). However, it is a fact that current languages change quickly, thanks to speakers' customs, knowledge, and social needs. In order to confirm this, we only need to follow very closely the growth of online abbreviation databases such as Acronym Finder, Acronyma, Abbreviations.com, Wiley InterScience, and All-acronyms.com, to cite but a few examples.

1.3 Abbreviation databanks: a solution to storing and managing initialisms

It can be said with certainty that the proliferation of initialisms, both in common and specialized discourse, has led to the need to store and manage them in a proper way; that is to say, by means of dictionaries and online databases. Apart from the case of Urology, Biomedicine is a good example for illustrating this phenomenon. Dannélls (2005) describes the current situation of the domain stating "Dictionaries are essential tools for understanding a language and are frequently used in many technical fields such as biomedicine where the vocabulary is quickly expanding. One known phenomenon in biomedical literature is the growth of new acronyms".

To date, there are several online databanks that have been created to store the great quantity of new initialisms. This kind of databanks can be classified into two categories, i.e., general and specialized.

A general initialism databank is a resource that contains both general and specialized discourse initialisms. In contrast, a specialized initialism databank is a resource created only for storing initialisms coming from a specific subject field.

'Acronym finder' is the most popular among the general initialism databanks. It was created in 1996 and has more than 4 million abbreviations. Abbreviations.com, which was created in 2001, is another well known databank. It classifies its abbreviations into 10 categories (Computing, Internet, Academic & Science, Miscellaneous, Medical, Business, Governmental, Community, Regional, and International), which are in turn subdivided into 132 subcategories. Finally, All-acronyms.com appeared in 2005. It

contains more than 600,000 abbreviations classified into 10 categories (Business, Education, General, Government, Locations, Medicine, Organizations, Phrases, and Science and Technology).

Specialized initialism databanks have also become relevant. At present, disciplines like Biomedicine have encouraged the development of good resources such as Acrophile, ADAM, Biomedical Abbreviation Server, Biomedical Acronym Resolver, and Acromed.

Those who create initialism databases, particularly the general ones, are simply interested in gathering the greatest possible quantity of initialisms without giving them a good lexicological/ terminological treatment. In fact, many data categories that might be helpful to different user profiles are not taken into account. Thus, entry records should include at least context, source, definition, domain, and when available, the equivalent initialisms in other languages.

1.4 Related work

Few comprehensive research works on initialisms from the Lexicology and Terminology point of view exist in Spanish.

First, from the domain of Lexicology, we highlight the work of Rodríguez (1981) "Análisis lingüístico de las siglas: especial referencia al español e inglés"1. This is a descriptive study that compares the linguistic features of initialisms with those of the common words of the language. The corpus employed in this study consisted of texts from newspapers, magazines and glossaries since it is a general discourse-oriented research.

Secondly, from the domain of Terminology, the work of Fijo (2003) "Las siglas en el lenguaje de la enfermería: análisis contrastivo inglés-español por medio de fichas terminológicas" stands out2. This is an empirical study of initialisms in the nursing domain. It describes all the aspects related to the creation, use, and translation of initialisms as terms belonging to the domain. It is also a bilingual study. It is based on a corpus that consisted of 65 initialisms in English and 58 in Spanish.

In short, it is clear that Spanish requires more descriptive and contrastive studies on lexical reduction units as well as the implementation of systems to recognize them in texts

2. Methodology

The following steps have been taken in order to carry out this research.

2.1 Initialism definition and typology. The following two tasks have been carried out: the study of the literature on the subject in Spanish, English, French and Catalan; and the terminology and lexicology view-points.

2.2 Corpus compilation. Both the Genomics and Environment subject field text corpus have been compiled. From these corpora the initialism corpus has then been built.

2.3 Linguistic description of initialisms. In order to carry out this step, the morphological, syntactic and semantic aspects of the initialism corpus have been analyzed.

2.4 Statistical description of initialisms. The sets of Genomics and Environment initialisms have been analyzed separately. This method allowed us to observe the characteristics of the initialisms in each subject field and to compare them later. In addition, two more corpora from Economics and Computer Science were contrasted with the previous ones.

2.5 Criteria for a Spanish initialism recognition system. Two different tasks have been done in order to perform this step. On the one hand, a study of the bibliography on current initialism recognition systems has been carried out. On the other hand, both the initialism formation rules and the initialism recognition patterns have been established.

3. Initialism – definition and typology

In the literature about lexical reduction units, there is a lack of consensus on the concept and typology of these units. In order to solve this problem, we have analyzed fifty-five different works in Spanish, English, French and Catalan. Checking all of them led us to establish a definition and typology in order to build and analyze our corpus.3

3.1 Definition of initialisms

An initialism is a lexical reduction unit made up of alphanumeric characters from a precedent lexical unit of syntagmatic structure. An initialism forms a sequence whose pronunciation can be spelled, syllabic or both; e.g.: PCR; TS; TEP; Grb2.4

3.2 Typology of initialisms

Initialisms can be classified into two different sets; i.e., proper and non-proper initialisms.

3.2.1 Proper initialism. Lexical reduction unit made up exclusively of initial characters of a precedent lexical unit of syntagmatic structure; e.g.: PCR (Polymerase chain reaction).

3.2.2 Non-proper initialism. Lexical reduction unit made up not only of initial characters but also of subsequent characters. A non-proper initialism is also a unit which has omitted basic parts of its expansion.

Non-proper initialisms can be classified into three subsets, i.e., typical non-proper initialisms, acronyms, and blends.

First, a typical non-proper initialism is a unit that uses or omits basic parts of its expansion and whose pronunciation can be spelled, syllabic or both; e.g.: Grb2 (Growth factor receptor-bound protein 2), SEF (superficie eficaz), etc.

Secondly, an acronym is a unit made up of several groups of characters from its expansion. Its pronunciation is exclusively syllabic; e.g.: HUGO (Human Genome Organization), ICONA (Instituto para la conservación de la naturaleza).

Thirdly, a blend is a unit similar to an acronym, but made up of the combination of two segments of the expansion. Its pronunciation is syllabic; e.g.: GeneBio (Geneva Bioinformatics), Agrimed (Agricultura mediterránea).

4. Corpus compilation

4.1 Genomics and Environment text corpus compilation. We have built our corpus from the corpus of the Institut Universitari de Lingüística Aplicada.5 The final output allowed us to build a Genomics text corpus of 1,028,964 words and an Environment text corpus of 1,027,741 words.6

4.2 Genomics and Environment initialism corpus compilation. We have employed a regular expression to query the text corpus in order to retrieve the initialism candidates. After manual refining, the Genomics and Environment initialism corpus amounted to 804 and 317 initialisms, respectively.

5. Linguistic description of initialisms

This study analyzes different qualitative aspects of the initialism corpus from a linguistic point of view. In order to do this, we have unified the Genomics and Environment initialism corpora. Then we have analyzed the morphological, syntactic and semantic features of each of them.

5.1 Morphological analysis

Regarding the morphological analysis, we have searched for cases of derivation and composition. In addition, we have determined the part of speech of the head word of the syntagm as well as its gender and number.

The only case of derivation found is the use of the adjectival suffix [-oso/a] in the initialism SIDA, i.e., sidoso.

As for composition we have registered the following cases:

5.1.1 prefix+initialism

pre-mARNS, pre-mARN, MegaYAC, antiVIH, retro-PCR, pre-ARNm

5.1.2 initialism+initialism

SARs-MARs, RT-PCR, NADP-NADPH, MVR-PCR, IR1/TR1, BCR/ABL

5.1.3 Part of speech of the expansion head word

All the head words of the initialism expansions are nouns.

5.1.4 Gender and number

With respect to gender, Spanish nouns fall into two categories, i.e., masculine and feminine. No variation was observed in the gender of the initialisms since they take it from the head word of the expansion they represent. On the contrary, we have found an interesting feature concerning the number of the initialism. In Spanish the duplication of the characters is the traditional mechanism for pluralizing; e.g.: CCOO (Comisiones Obreras). However, we have observed initialisms in our corpus that have been pluralized according to the English mechanism, i.e. by adding the character 's' at the end of the initialism like DNAsas and rARNs. It seems that the general trend in specialized discourse in Spanish is the use of the English pluralization mechanism.

5.2 Syntactic analysis

For this analysis we have studied two different aspects: a) the type of syntagm, and b) the units that usually co-occur with the initialisms.

5.2.1 Type of syntagm

The analysis of the corpus has shown 1,121 nominal syntagms.

5.2.2 Lexical combination

The corpus has shown the presence of different combinations of lexical units with initialisms, as can be seen from the following list.

- [N+initialism]; e.g.: tecnología PCR, marcador RFLP, secuencias VIH, gen HLA.

- [N+Prep+initialism]; e.g.: oligonucleótidos de PCR, protocolos de PCR, análisis con RFLP, portadores de FQ.

- [Initialism+Adj]; e.g.: PCR multiplex, RFLP dialélico, VIH residuales, MHC humano.

- [V+initialism]; e.g.: los autores utilizan PCR…, se denomina RFLP…, se denominan VNTR…, para amplificar VNTR…, etc.

5.2.3 Semantic analysis

In our study, the semantic analysis comprises several aspects; however, here we focus on the formal variation, i.e., the co-occurrence of initialisms and their expansions.

- Formal variation

We consider every initialism a synonym of an existent syntagmatic unit. We call this phenomenon formal variation; e.g.: PGH (Proyecto Genoma Humano), VHC (Virus de la hepatitis C), LINEs (Long interspersed nuclear elements), etc.

From all the above, it can be seen that to a greater or lesser degree, initialisms evince the same linguistic features as words from the point of view of morphology, syntax and semantics. In addition, when an initialism adopts composition and derivation elements and changes from uppercase to lowercase characters like sida, it becomes a lexicalized initialism; that is to say a word.

6. Statistical description of initialisms

As we have said before, the text corpus of Genomics consisted of 1,028,964 words whereas the Environment corpus consisted of 1,027,741 words.

6.1 Descriptive statistics of Genomics initialisms

The following are the aspects taken into account in the corpus analysis: 1) initialism quantity, 2) initialism occurrences, 3) initialism-expansion co-occurrences, 4) general discourse related initialisms (institution names), 5) specialized discourse related initialisms (terms), and 6) the quantity of English initialisms versus the quantity of Spanish initialisms. These are the results of the analysis.

- 804 different initialisms that represent 11,314 occurrences (1.09% of the text corpus).

- 467 initialisms (58%) show their expansion (formal variation) in the corpus.

- 71 initialisms (9%) belong to general discourse (institution names) whereas 733 (91%) belong to specialized discourse (terms).

- 418 initialisms (52%) are proper whereas 384 (48%) are non-proper.

- 652 initialisms (82%) represent expansions originally created in English whereas 147 (18%) represent their expansions in Spanish.

In conclusion, it can be said that the total percentage of initialisms in this corpus seems to be high (more than 11,000 occurrences). In addition, there is a high percentage of initialisms representing terms (91%). Likewise, 82% of initialisms represent expansions created in English.

6.2 Descriptive statistics of Environment initialisms

- 317 different initialisms that represent 1,593 occurrences (0.15% of the text corpus).

- 194 initialisms (61%) show their expansion in the corpus.

- 166 initialisms (52%) belong to general discourse (institution names) whereas 151 (48%) belong to specialized discourse (terms).

- 219 initialisms (69%) are proper initialisms whereas 98 (31%) are non-proper.

- 94 initialisms (30%) represent expansions created originally in English whereas 194 (61%) represent their expansions in Spanish.7

In conclusion, it can be said that the total percentage of initialisms in this corpus seems to be low, only 1,593 occurrences of initialisms in more than 1 million words. In addition, about 50% of initialisms correspond to institution names, and 30% belong to expansions created originally in English.

6.3 Genomics and Environment comparative descriptive statistics

A comparison between the two corpora shows that the quantity of initialisms is higher in the Genomics than in the Environment corpus (Fig. 1). However, formal variation percentages are similar in both corpora (Fig. 2). Initialisms represent more terms in the Genomics than in the Environment subject field corpus (Fig. 3). Finally, Genomics shows a higher percentage of non-proper initialisms and English initialisms than does Environment. Conversely, Environment shows the highest percentages of proper initialisms and Spanish initialisms (Fig. 4).

The Genomics and Environment corpora were compared to two more corpora, i.e., a Computer science corpus (1,027,995 words) and an Economics corpus (1,029,172 words). For the analysis of these corpora we have taken into account three parameters only: 1) the quantity of initialism occurrences, 2) initialism-expansion co-occurrences, and 3) general and specialized discourse related initialisms.

6.4 Descriptive statistics of the Computer science corpus

- 876 different initialisms that represent 6,130 occurrences (0.6% of the text corpus).

- 568 initialisms (65%) show their expansion in the corpus.

- 59 initialisms (7%) belong to general discourse whereas 817 (93%) belong to specialized discourse.

6.5 Descriptive statistics of Economics

- 244 different initialisms that represent 1,708 occurrences (0.2% of the text corpus).

- 141 initialisms (57%) present their expansion in the corpus.

- 136 initialisms (56%) belong to general discourse whereas 108 (44%) belong to specialized discourse.

6.6 Genomics and Environment corpora versus Economics and Computer Science corpora

A comparison between the four corpora has shown that the highest percentage of initialisms appears in the Genomics and Computer Science corpora, respectively (Fig. 5). Likewise, these two corpora have shown the highest frequency of specialized initialisms. With respect to formal variation, the Computer Science and Environment corpora present the highest frequency. Finally, the Economics and Environment domains present the highest quantity of general discourse initialisms, basically institutional names (Fig. 6).

7. Criteria for a Spanish initialism recognition system

In the domain literature we have found two main motivations for developing an initialism recognition system. One is the need for automatic update of initialism databases, and the other is the need to facilitate information retrieval in a given knowledge area.

The initialism recognition process comprises three different steps: 1) initialism identification; 2) initialism-expansion pair identification, and 3) initialism disambiguation.

First, a set of initialism formation rules must be established. In order to perform this step, we have compared the analysis of our corpus with the works of Taghva & Gilbreth, 1999; Pustejovsky, 2001; Ananiadou, 2002; Larkey, 2002; Zahariev, 2004, and Dannélls, 2005. Secondly, a set of initialism-expansion pair recognition patterns must be determined. Thirdly, in order to disambiguate the senses of initialisms, some systems use two techniques, namely the inclusion of abbreviation dictionaries or the analysis of initialism-expansion in context. Here we shall focus on the first and second aspects.

In order to recognize an initialism candidate it is necessary to know its formation rules and its recognition patterns.

7.1 Initialism formation rules. There are two types of initialism formation rules, i.e., basic and complementary.

7.1.1 Basic rules of initialism formation

- More than 2 and less than 9 characters

- Maximum 2 words

- A minimum of one uppercase character

- First character alphanumeric

- Omission of symbols such as ; : ? !

7.1.2 Complementary rules for initialism formation

7.2 Initialism recognition patterns.There are two types of patterns for recognizing initialisms, i.e., initialism candidate recognition patterns and initialism-expansion pair recognition patterns.

7.2.1 Initialism candidate recognition patterns

Based on the analysis of our corpus and the study of Larkey et al. (2000), we have established the following patterns:

- An initialism may show between 2 and 9 uppercase characters. It may contain both full stops and plural marks; e.g.: U.S.A, U.S.A.'s. This pattern is represented as follows: (U {sep})2-9S, where

U= Uppercase

{sep}= full stop or full stop followed by a space

2-9= rank of characters

S= plural mark (Not all initialisms show them. They are very frequent in initialisms created in English).
- An initialism may show from 2 to 9 characters and a plural mark as well; e.g.: USA, USA's. This pattern is represented as follows: U2-9S, where

U= Uppercase

2-9= rank of characters

S= plural mark (Not all initialisms show them. They are very frequent in initialisms created in English).
- An initialism may show zero or one or more character occurrences in uppercase followed by a number between 1 and 9 and one or more occurrences of uppercase characters; e.g.: 3D, 3-D, I3R. This pattern is represented as follows: U*{dig}U+, where

U= Uppercase

*= 0 or more occurrences

{dig}= number between 1 and 9, followed by a dash (optional)

+ = one or more occurrences of a character.
- An initialism may contain one or more lowercase characters followed by one or more uppercase characters; e.g.: DoD. This pattern is represented as follows: U+L+U+, where

U= Uppercase

*= 0 or more occurrences

+ = one or more occurrences of a character

L= Lowercase.
- An initialism may be formed from one or more uppercase characters; e.g.: AFL-CIO. This pattern is represented as follows: U+[/-]U+, where

U= Uppercase

+ = one or more occurrences of a character

[/-]= separation character such as slash or dash.

7.2.2 Initialism-expansion pair recognition patterns

Based on the works of Larkey et al., 2000; Pustejovsky, 2001; Schwartz & Hearst, 2003; Nenadic, 2002; Adar, 2004; Nadeau & Turney, 2005; and Dannélls, 2006; as well as on the analysis of our corpus, we have established a list of the 10 most frequent initialism-expansion pair recognition patterns, which can be seen in table 1.

#	Pattern	Genomics	Environment	Pattern found in other works Yes/No
		Frequency	Frequency	
1	EXPANSION (INITIALISM)	295	192	Yes
2	INITIALISM (EXPANSION)	102	52	Yes
3	INITIALISM ("EXPANSION")	13	1	No
4	INITIALISM, EXPANSION	12	6	No
5	"EXPANSION" (INITIALISM)	9	8	Yes
6	(INITIALISM, EXPANSION)	9	2	No
7	EXPANSION o INITIALISM	8	-	Yes
8	EXPANSION (INITIALISM,)	6	-	No
9	(EXPANSION, INITIALISM)	5	1	No
10	EXPANSION, INITIALISM	4	4	No

Table 1. The most frequent expansion-initialism recognition patterns

Both the analysis of previous work and our corpus have shown that the most frequent pattern is **"Expansion (Initialism)".**

In addition, for recognizing initialisms in Spanish texts it is necessary to bear in mind a mixed set of patterns since the following cases frequently occur:

- Both the initialism and the expansion are in Spanish.

- The initialism is in a foreign language but its expansion is translated into Spanish, and, conversely,

- The initialism is translated into Spanish, but its expansion remains in the foreign language.

Conclusions

The study of initialisms in the Genomics and Environment subject fields allowed us to come to the following conclusions:

1. Initialisms are a generalized phenomenon since they are found at present in many languages and domains.

2. Initialisms are considered a lexical reduction phenomenon. Today it is easy to find specialized texts full of them, a clear evidence of terminological variation. Initialisms can be lexical and semantic variants since they are used as synonyms of their expansions. Initialisms are pragmatic variants and they are supposed to facilitate the expert's reading.

3. Initialisms have attracted wide interest on the part of lexicographers, translators, linguists, technical writers, and subject field experts (mainly from Biomedicine). However, despite the fact that initialisms represent the shortening of equivalent terms, no attention has been paid to the description and analysis of them from the terminology point of view.

4. Although many authors have studied initialisms before, their research has been focused on English in the vast majority of the cases. Consequently, there is a lack of description and analysis in other languages such as Spanish. In fact only one work exists which has dealt with this subject under such a perspective, viz. Fijo (2003). In addition, all the previous research has shown inconsistencies in the definition and typology of initialisms.

5. The study of initialisms is relevant in Terminology since they appear frequently in specialized texts. Moreover, apart from being important for the above mentioned domains, initialisms are especially

important for Natural Language Processing, Artificial Intelligence, Data Mining, Information Retrieval, and Machine Translation purposes.

6. Although many authors disagree with the generalized use of initialisms, current language shows that it is a growing phenomenon.

7. At present, many online abbreviation databases function as large repositories. However, neither lexicographical nor terminographical methods have been applied on them.

8. Some domains are more influenced by initialisms than others. For example, Genomics and Computer Science texts show higher initialism occurrence than Environment and Economics texts. Perhaps this may be ascribed to the use of high technology and the advanced research techniques used in the former domains.

9. Currently, research on initialisms shows two trends: theoretical description and automatic identification and disambiguation from texts.

10. Initialisms are supposed to facilitate the expert's reading. However, in the case of laymen, initialisms become a problem due to their "opacity" when they do not co-occur with their expansions. Likewise, initialisms make the tasks of Natural Language Processing difficult.

11. Up to now, Biomedicine is the domain that has devoted the most research efforts to developing initialism recognition systems.

12. Almost all recognition systems have been implemented for the English language. Currently, only one system has been developed for recognizing initialisms in medical texts in Swedish (cf. Dannélls, 2005; 2006).

Acknowledgements

This paper is part of the research carried out in the PhD Thesis "Descripción y análisis de las siglas en el discurso especializado de Genoma humano y Medio ambiente", supervised by Professor Maria Teresa Cabré. This work has been supported by the Generalitat de Catalunya Research Grant (2004 FI 00400), and the Research Project "Texterm II. Fundamentos, estrategias y herramientas para el procesamiento y extracción automáticos de información especializada" (MCYT, BFF2003-2111, 2003-2006), carried out at the Institut Universitari de Lingüística Aplicada (IULA) of the Universitat Pompeu Fabra.

Notes

1PhD thesis

2PhD thesis

3Cf. Annex 1.

4We consider as lexical reduction units (or abbreviations) every phenomenon of lexical shortening.

5Cf http://bwananet.iula.upf.edu/indexen.htm

69% of initialisms represent expansions in other languages such as French, German, Catalan, etc.

References (A selection)

ABBREVIATIONS.COM [on line] http://www.abbreviations.com/about.asp

ACRONYMA [on line]. http://www.acronyma.com/

ACRONYM FINDER [on line]. http://www.acronymfinder.com/

ADAR, E. (2004). SaRAD: A simple and Robust Abbreviation Dictionary. Bioinformatics, 20 (4) 2004. 527-533.

ALL-ACRONYMS.COM [on line]. http://www.all-acronyms.com

ANANIADOU, S. et al. (2002). Term-based Literature Mining from Biomedical Texts. [on line]. http://www.pdg.cnb.uam.es/BioLink/Ananiadou.doc

BIOMEDICAL ACRONYM RESOLVER ARGH. [on line]. http://invention.swmed.edu/argh/

BLOOM, D. A. (2000). Acronyms, abbreviations and initialisms. BJU International, 86. 1-6.

BRADLEY, J. (2004). The Acronym addiction. Texas Heart Institute Journal, 31 (1). 108-109.

CABRÉ, M. T. (1999). La terminología: representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos. Barcelona: Institut Universitari de Lingüística Aplicada.

CABRÉ, M. T. (2003). Theories of Terminology: their description, prescription and explanation. Terminology, 9 (2). 163-200.

CHANG, J. et al. (2002). Creating an Online Dictionary of Abbreviations from MEDLINE [on line]. http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=12386112

CHENG, T. (1997). Non-English acronyms must be explained in their native languages. International Journal of Cardiology, 61. 199.

CHENG, T. (2002a). Acronyms must be defined. Atherosclerosis, 165. 383.

CHENG, T. (2002b). Every acronym should be defined when it first appears in a publication. Circulation, 106. 134.

CHENG, T. (2005). Celestial acronyms. International Journal of Cardiology, 101. 307-308.

DANNÉLLS, D. (2005). Recognizing Swedish acronyms and their definitions in biomedical literature. [on line]. http://www.cling.gu.se/~cl2ddoyt/acronyms/report.pdf

DANNÉLLS, D. (2006). Automatic Acronym Recognition. Proceedings of the 11th conference on European chapter of the Association for Computational Linguistics. [on line]. http://www.cling.gu.se/~cl2ddoyt/pub/automatic.pdf

FALLOWFIELD, L. (2002). Acronymic trials: the good, the bad, and the coercive. The Lancet, 360. 1622.

FIJO, M. I. (2003). Las siglas en el lenguaje de la enfermería: análisis contrastivo inglés-español por medio de fichas terminológicas. PhD Thesis. Sevilla: Departamento de Humanidades, Universidad Pablo de Olavide.

FRED, H. (2003). Acronymesis. The Exploding Misuse of Acronyms. Texas Heart Institute Journal, 30 (4). 255-257.

GREEN, W. (1990). Abbs. in Js. Canadian Medical Association Journal, 142 (4). 287.

LADER, E. (2002). Acronym mania. The Lancet, 160. 576.

ISAACS, D. (2007). Acronymophilia: an update. ADC, 83. 517-518.

JACK, D. (2003). The cardiology SCANDAL. The Lancet, 361. 538.

LARKEY, L. et al. (2000). Acrophile: An Automated Acronym Extractor and Server. [on line]. http://delivery.acm.org/10.1145/340000/336664/p205-larkey.pdf?key1=336664&key2=7455896901&coll=GUIDE&dl=GUIDE&CFID=28595879&CFTOKEN=50021223

MORGAN, P. (1985). A quick look at medical abbreviations. Canadian Medical Association Journal, 32. 897.

NADEAU, D.; Turney, P. (2005). A Supervised Learning Approach to Acronym Identification. [on line]. http://iit-iti.nrc-cnrc.gc.ca/iit-publications-iti/docs/NRC-48121.pdf

NENADIC, G. et al. (2002). Automatic Discovery of Term Similarities Using Pattern Mining. [on line]. http://acl.ldc.upenn.edu/coling2002/workshops/data/w05/w05-08.pdf

PUSTEJOVSKY, J. et al. (2001). Linguistic Knowledge Extraction from Medline: Automatic Construction of an Acronym Database. http://www.medstract.org/papers/bioinformatics.pdf

RODRÍGUEZ, F. (1981). Análisis lingüístico de las siglas: especial referencia al español e inglés. PhD Thesis. Salamanca: Facultad de Filología, Universidad de Salamanca.

ROWE, R. (2003). Abbreviation Mania and Acronymical Madness, DDT 8 (16). 732-733.

SCHWARTZ, A.; Hearst, M. (2003). A Simple Algorithm for identifying Abbreviation Definitions in Biomedical Text. [on line] http://biotext.berkeley.edu/papers/psb03.pdf

TAGHVA, K.; Gilbreth, J. (1999). Recognizing acronyms and their definitions. International Journal on Document Analysis and Recognition. 191-198. [on line] https://troia.upf.edu/http/www.springerlink.com/content/u6c9ymd1v8jflerh/fulltext.pdf

VANDAELE, S.; Pageau, M. (2006). Dynamique discursive et traduction des signes abréviatifs en biomédecine. Équivalences, 33 (1-2). 165-190.

WALLING, H. (2001). When will the MEK inherit the ERK? Acronym alphabet soup. TRENDS in Pharmacological Sciences, 22 (1). 14.

WILEY INTERSCIENCE [online]. http://www3.interscience.wiley.com/cgi bin/home?CRETRY=1&SRETRY=0

ZAHARIEV, M. (2004). A(Acronyms). PhD Thesis. Burnaby: School of Computing Science, Simon Fraser University. [on line]. http://www.cs.sfu.ca/~manuelz/personal/p/f.pdf

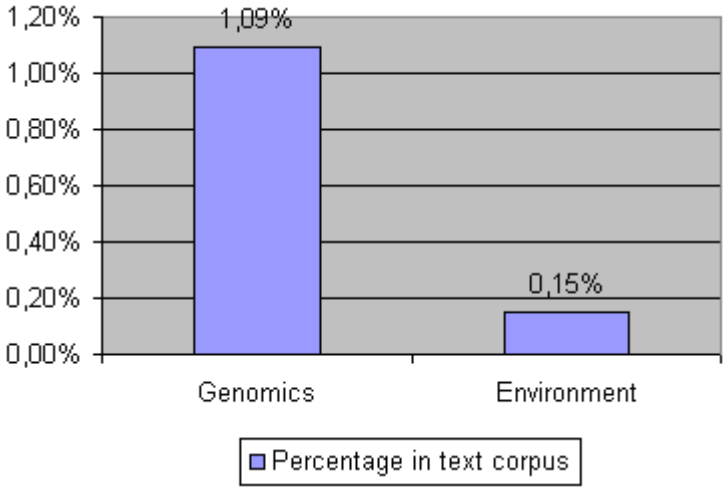


Fig. 1. Initialism percentages in text corpus

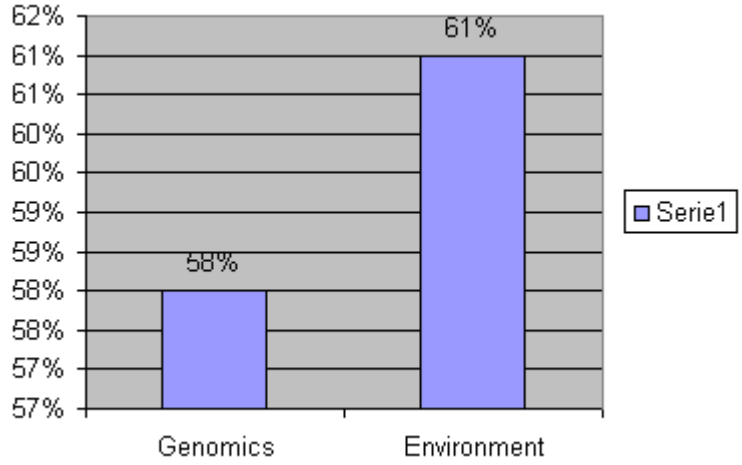


Fig. 2. Formal variation or expansion-initialism co-occurrence percentages

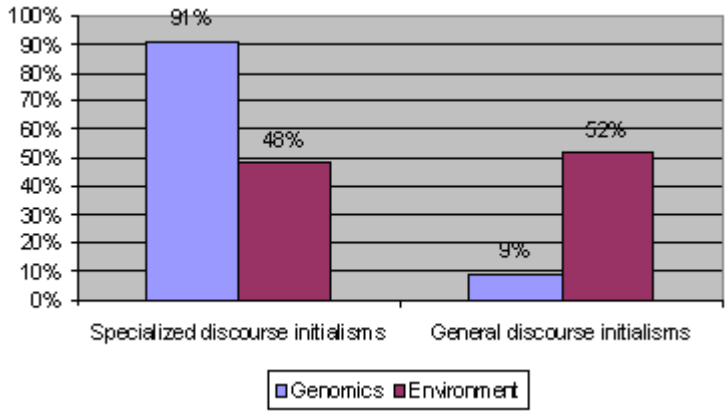


Fig. 3. Specialized discourse initialisms vs General discourse initialisms

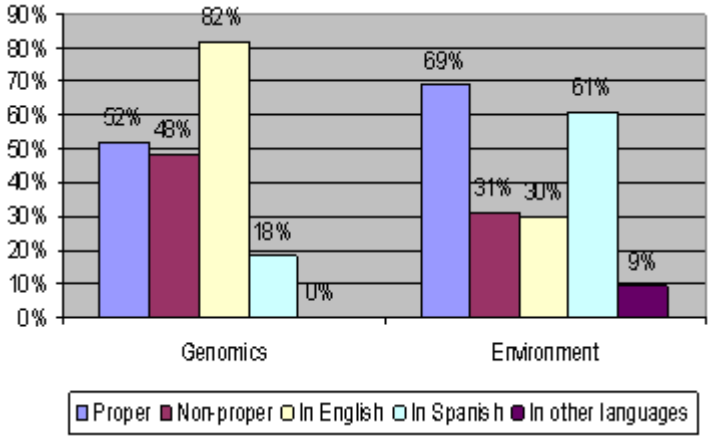


Fig. 4. Percentages by type of initialisms and languages in which initialisms have been created

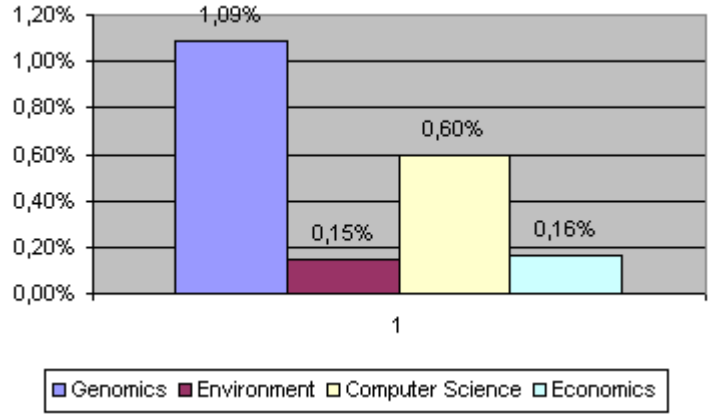


Fig. 5. Initialism percentages in text corpora

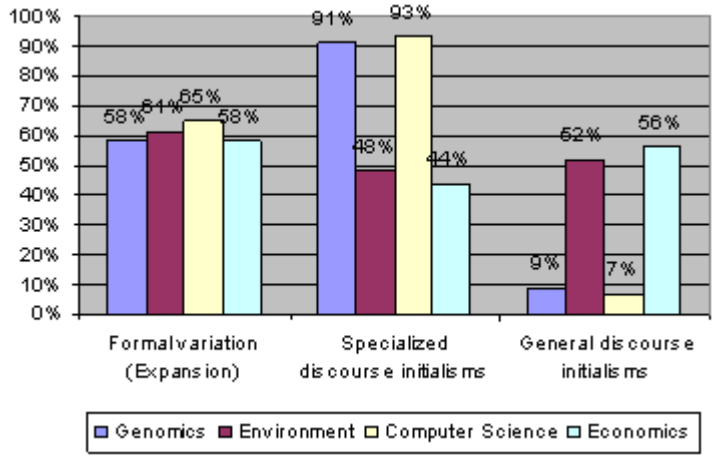


Fig. 6. Initialism percentages for formal variation, specialized discourse and general discourse in the four corpora

Group 1
Authors that do not make a distinction between initialisms and acronyms:

1. Burnett (1994)
2. Colás (1994)
3. Capó & Veiga (1997)
4. Estopà (2000)
5. Larkey (2000)
6. Lázaro Carreter (1990)
7. Matthews (1997)
8. Mejía (1980)
9. Mestres i Serra (1985, 1995)
10. Mitterand (1986)
11. Mounin (1982)

Group 2
Authors that support that an acronym differentiates from an initialism by its pronunciation or formation process.

Trend 1: Syllabic pronunciation

1. Algeo (1991; 2003)
2. Bastons & Font (2001)
3. Brusaw (1987)
4. Bussmann (1996)
5. Calvet (1980)
6. Crystal (2003)
7. Diccionario Le trésor de la langue fraçaise informatisé (2002)
8. Diccionario Merriam-Webster (2003)
9. Gran diccionari de la llengua catalana (2003)
10. Gehénot (1990)
11. Huddleston *et al.*(2002)
12. López Rúa (2000)
13. Losson (1990)
14. McArthur (2003)
15. Mossman (1992)
16. ISO 1087-1 (2000)/ ISO 12620 (1999)
17. Percebois (2001)
18. Pérez Saldanya (1998)
19. Vandaele & Pageau (2006)
20. Zolondek (1991)

Trend 2: Joint of opposite extrems of two words

1. Alvar & Miró (1983)
2. Cabré (1993)
3. Cardero (2002)
4. Casado Velarde (1985)
5. Diccionario Larousse/Termcat (1992)
6. Fijo (2003)
7. Martínez de Sousa (1984)

Group 3
Authors that consider acronyms as a type of initialism:

1. Abreu (1997)
2. Alcaraz (2003)
3. Alcaraz & Martínez (1997)
4. Arntz & Picht (1995)
5. Bezos (2007)
6. Cardona (1991)
7. Cerdà (1986)
8. Diccionario DRAE (2001)
9. Diccionario Le Nouveau Petit Robert (2001)
10. Dubois (1994)
11. Fernández (2002)
12. Gouvernement du Québec (2002)
13. Grammaire reverso.net (2000)
14. Maldonado (2002)
15. Mestres & Guillén (2001)
16. Nakos (1990)
17. Quirk (1985)
18. Rodríguez González (1981)

Klaus-Dirk Schmitz
Institute for Translation and Multilingual Communication
Cologne University of Applied Sciences
Cologne (Germany)

COMMENTS ON "DESCRIPTION AND ANALYSIS OF INITIALISMS IN THE GENOMICS AND ENVIRONMENT SUBJECT FIELDS"

Within the framework of the session "New Voices in Terminology and Future Research Directions" at the 2007 LSP Conference in Hamburg (Germany), John Jairo Giraldo Ortiz from Medellin (Colombia) presented a research paper with the title "Description and Analysis of Initialisms in the Genomics and Environment Subject Fields". The objective of this paper is twofold: it analyses the initialisms in LSP texts of the domain of genomics and environment by means of linguistic and statistical methods, and it tries to establish patterns for the automatic recognition of candidates for the creation of abbreviation dictionaries. The approaches and results described are part of the author's PhD Thesis and a research project carried out at the University Pompeu Fabra in Barcelona (Spain).

John Jairo Giraldo Ortiz starts his paper with a short discussion of the general background of his research topic. It is followed by a concise description of the underlying methodology and a short definition and typology section. The main part of his research is reflected in sections 4 to 6, where the corpus constitution as well as the corpus qualitative and quantitative analysis are described in detail by a large number of Spanish examples and statistical figures. The second objective of Giraldo's research paper is described in section 7, in which – based on the results of the previous investigations – he develops criteria for an (automatic) recognition of Spanish initialisms. A conclusion and a selected bibliography complete the paper.

The topic of Giraldo's research paper is innovative. Although there exist many scientific papers dealing with abbreviated forms of terms and initialisms, the vast majority concentrates on the English language and focuses more on linguistic than on terminological aspects. The selection of the two subject fields, genomics and environment, is adequate and diverse enough for the intended research, and the results show that LSP texts in both domains behave differently in the field of initialisms. The analysis is based on sound statistical data; both Spanish corpora – as well as the "control corpora" with computer science and economy textual material – contain more than 1 million words each.

The corpus qualitative and quantitative analysis is done in a very profound way. The morphological and syntactic patterns and combinations are described and differentiated in detail and highlighted with statistical data from the corpus analysis. The semantic analysis is less deep, but this is understandable given that the focus of the research is mainly on an automatic initialism recognition system.

The section related to the Spanish initialism recognition system shows that Giraldo has scrutinized the main international approaches for term and initialism recognition as well as for initialism-expansion identification using text corpora. Giraldo's recognition patterns both for initialisms and for initialism-expansion pairs are well elaborated and based on his own corpus analysis material. Only in future research can it be ascertained if an automatic initialism recognition system based on his patterns and rules will detect all occurrences (recall and precision) in subject specific textual material (in all domains).

Although Giraldo's starting approach does not take into account the typology of abbreviated forms of terms defined in (e.g. German) national and international standards, his concentration on initialisms is comprehensible in the framework of his objectives. His research work, taking into account existing approaches and based on an excellent theoretical and methodological background as well as on sound statistical data, provides a major contribution to the scientific community.