

Some Considerations for the Semantic Web

María Elena Franco Carcedo

ICUAP, CA RI-FCC-BUAP, Ciudad Universitaria,
72570 Puebla, Mexico
fcarcedo@hotmail.com

Abstract. Here are some considerations about the management and level of verbal language domain of semantic web potential users, their characteristics, most frequent errors and consequences, in order to be took into account for a better performance of the semantic Web.

Keywords: Semantic Web, verbal language.

1 Introduction

This paper has been constructed on the basis of vast experience gained through several years on the characterization of the academic language (record/register). Moreover, this research is the result of a deeper study on postgraduate careers of a public university (BUAP), which let us to detect the most frequent mistakes and transgressions on the academic language.

We will start this discussion with two questions: 1) What is the meaning of those terms that we hear and use, but whose conceptualization or definition will state difficulties of some kind? In academia, especially in recent years, there are some terms that are frequently used within institutional development plans. Terms such as: pedagogy, didactics, methodology, teaching techniques, teaching process learning, etc are used as a prelude of academic excellence and resource optimization (human resources and materials). And here becomes the second question: Does not imply knowledge any kind of specific knowledge, a specific language, a certain jargon? If we accept this assertion as a claim we may infer that by acquiring a scientific knowledge of language (knowledge instrument¹) we can facilitate the language knowledge and application of the same. But the real question is the following: Does the use of a

¹ There is a theoretical discussion about the status of the studies dealing with the language, from *grammar* to the diverse branches, lines and research. By centuries it was considered the *grammar* and *rhetoric* and *oratory* as complete arts. However for the end of last century, with F. de Saussure as a benchmark, it was understood as a *science*. In this paper we will not go deeper in order to argue about one or other viewpoint.

² We would argue whether or not the natural language is the knowledge instrument or the knowledge itself with respect to its role as reality creator, in terms of the *rational speech* or thought. *How do we know? How do we learn the reality?* By means of which mechanism or procedure we “know” and “know the world”.

semantic Web require some user verbal skills or at least to advise the lack of these skills?

2 Considerations

Errors or deficiencies affect at different levels the following issues: *intelligibility* (Rules 1, 2, 3 of verbal system); *the thought logic* (Rules 3, 4, 5); *communication* (null or minimal noise in the process of encoding-decoding); *speech fluidity and precision* (speech alludes to both phrasing and reasoning). Natural language is a system of verbal symbols (sorted set of functioning rules of constituent items). The academic subsystem (record/register) and jargon (characteristic of the different areas of knowledge) is characterized by the following rules.

2.1 Phonetic-phonological and spelling rules

- Signs Graphics: punctuation and auxiliary signs.
- Spelling: diacritics, punctuation, accentuation, ortho-semantic problems.

2.2 Lexical-semantics rules

Lexical (3rd unit of language, 1st significant unit): univocity, precision, accuracy, conciseness, versatility and richness vs. ambiguity, categorization error, polysemy, synonymy, homonymy, homophony, paronymie, imprecision, vagueness, semantic phrases, poverty, repetition, ignorance, confusion.

2.3 Morphosyntactic rules

- Constructive (5th unit of language, 1st logical unit): logic, consistency, sequential, cohesion, consistency, clarity, simplicity vs. illogical (juxtaposed and copulative; alteration of conjunctive nexus, etc.) assumptions, fractures, anacolutha, truncated sentences, inconsistency (especially subject-predicate), confusion (especially in terms of managing the elements of the compound sentence) , scavenging, and Pseudo-cultism (ex, abuse of “el cual” - which); incorrect temporal correlations, and so on.
- Concordance, among others.

2.4 Encoding rules

- Encoding: denotative, which corresponds to the formal record/register (academic, scientific, technical) vs. subsystem of use, characterized by a

tendency of connotative encoding, the standard or colloquial register, oral, with semantic shifts or meanings tagged by chrono- and geo- sociolects –the use of archaisms (a word or form of speech no longer in common use); or only functional in some areas or among certain social groups – altered prepositional regimen, etc.

2.5 Logical-stylistic rules

- Structure (syntax of the compound sentence, what relationships are established: cause-consequence, principle-purpose, conditional potential, hypothesis, etc.); the ideas exposition and their linking; structure consistent with the basic genre on introduction (definition, description, narrative, argumentation), development and conclusion.
- Functional: the event affects the purpose, theme (or its treatment), target and record/register (what, for whom, for what purpose, how); informative function, logic, data or knowledge transfer, assertive or speculative, formalized as article, essay, notes, summary; without subjective, intimate or emotional levels.
- Style: neutral and objective tone; author plural or third impersonal; without dialect marks; rigorous, light, clear and simple.

We are particularly interested on those problems related with the *word* unit. Therefore, we will not stress on unacceptable underlying assumptions, stemmed phrases, anacolutha and diverse fractures, erroneous punctuation, with the resulting disruption of the correct decoding or in cases of disagreement, but we will focus our attention on aspects related with Rules 1 and 2: **Phonetic-phonological and spelling rules; Lexical-semantics rules: transgressions and most frequent errors** and we will present some examples.

- Vocabulary deficiencies and use meanings³
- Confusion/alteration of logical-grammatical category
- Spelling errors that affect meaning (written accent mark: *íntegro-integro-integró; gráfica-gráfica*, etc.).
- Problems with homophones, homonyms, paronymie and polysemy⁴
- Abuse of some Spanish verbs when defining and writing, such as: *ser*(being), *estar*(to be), *tener*(to have), *poder*(to be able), *hacer*(to do).
- False friends from the English language (Due to the users read literature in English: *remover* (remove) instead of *eliminar*, *interface* or *interfase* (interface) instead of *interfaz*, etc.).
- Abuse of lexical and constructive Anglicism and the problems they cause

³ In order to simplify and due to we are interested on the determination of the dialectic phenomena, we create the concept **of use** sub-system. We will focus on transgressions with respect to the academia norms that rule this record/register.

⁴ In both senses, strict and soft, i.e., we include here the meaning plurality.

- Misuse of the gerund (English gerund)
- Forms of 'passive English' (construction of the auxiliary 'to be' (*ser*) rather than reflected passive)

3 Experimental results

In this section we present an evaluation of a written test carried out on 135 students of different postgraduate programs on science and technology. We asked them to define different words. Even if the complete list was conformed by 25 terms, we only show the obtained results with a subset of this set. In summary, we have detected 13 problems, which are described as follows.

1. **Ignorance of the term meaning (null answer).** Surprisingly, not one word were answered in blank (null answer), even on terms such as *vaso*(glass), *sección*(section), *adolescente*(teenager), where answer percentage rates were 1,48%, 2,22% and 9,62%, respectively. Percentages rates reach values such as 67,4% (*acerbo*), 65,92% (*asechar*) and 50,37% (*sito*).
 - The pair *acerbo-acerbo*, had 91 and 51 null answers (67,4% and 37,7% respectively), from which only 2 and 46 were partially correct (2,2% and 34%).
 - The pair *adolescente-adolescente*, had 34 and 13 null answers (25,1% and 9,6%), from which 37 and 69 were correct (from this 37, only 5 were totally correct, i.e., those that gave more than one meaning; in the second case, i.e., *adolescente*, there is only one meaning which may be expressed in different manners).
 - The pair *baso-bazo*, had 66 and 34 null answers (44,8% and 25,1%), from which 29 and 93 were correct (partially correct in general, since in major cases there were given only one meaning) with percentages of 21,4% and 68,8%, respectively.
2. Confusion of logical-grammatical category within the term meaning.
 1. Simple.
 - **Term:** *soluble* (adj.) **Definition given:** 'disolver' (verb);
 - **Term:** *cesión* (noun) **Definition given:** 'que se termina o *sesa* [*sic*];
 - **Term:** *suspendido* **Definition given:** '*cuando* algo a [*sic*] terminado ó [*sic*] concluido'
 2. Half confusion (one category inside and one category outside).
 - **Term:** *acceso* (sust.) **Definition given:** *entrada* (sust.), *viable* (adj.);
 - **Term:** *sesión* (sust.) **Definition given:** 'congregarse personas (verbo), reunión (sust.)'

3. Confusion with its homophones (diacritical spelling) or paronymie. The value ranges from 0% for *vaso* (*baso-bazo*) to 17,7%, 17,3% and 14,8% for *asechar*, *adolescente*, *rebelar*, which are defined by their pair *acechar*, *adolescente*, *revelar*, respectively.
1. Homophone category error. *acerbo* (adj.): *conjunto de bienes comunes* (corresponde a *acervo*, sust.)
2. Outside of its category [Example. *Acerbo* (adj.): *tener conocimientos* (they mislead with *acervo*, sust., but they defined it as verb)
3. Half confusion (one category inside and one category outside). *acerbo*: 'acumular, conjunto de cosas'
4. **Answer outside of the semantic field.** Unbelievable at first glance, among the highest percentages it was found the term *vaya* (exclamation and verb) with the 77,77%, followed by *haciendo*, with 69,62%, *vario*, with 57,77%. Examples of wrong field:
 - **Term:** *adolescente* [person] **Definition given:** 'etapa'; 'edad' [cosa];
 - **Term:** *asciendo* (present tense verb, act) **Definition given:** 'posición en que uno queda' [resultado de un acto];
 - **Term:** *revelar*, **Definition given:** 'acción química [sic] de la fotografía'
5. **Definition by means of:**
 - Temporality: *es cuando* (ej. *estática*: cuando **una persona** no se mueve)
 - Mode: *es como...* (it is like...) (ex. *occisa*: *es como estar muerta*)
 - Location: *es donde...* (is where...) (ex. *jardín*: *donde juegan los niños*)
 - Function: *sirve para...* (it is to...) (ex. *lenguaje*: *sirve para comunicarse*)
6. **To classify instead of defining.** Very often they use to annotate the infinite without definition (nor meaning) in the particular case of conjugated verbs. Example. *izo*: verbo *izar*; *vacilo*, 'de vacilar'.
7. **To complete instead of defining.** (example: *acervo*: *bibliográfico*; *sucesión*: *presidencial*)
8. **The definition was too much...**
 - Extended, wide (ex. *es una cosa...*; *es algo...*)
 - Restricted (ex. *acervo*: *conjunto de palabras*)
 - Fuzzy, imprecise (ex. *vacilo*: *pertenece a la biología*)
9. **Meanings or homonymy omission** (we are only interested on those answers that show significant plurality but not the classification of the term meaning of homonym term). This is considered "partial correct answer". In *rebelar*, it was obtained a 41,1%; in *revelar* the 46,6%; in *ascender* the 48,8%, i.e., the majority.
10. **Figurative sense, without the denotative.** *árido*: 'falto de amenidad'
11. **Semantic shift.** *árido*: 'molesto'
12. **Subsystem of use or colloquial.** *vacilar*: 'echar relajo'
13. **Answer without meaning.** *arriar*: 'arroyar' [por *arrollar*]

In summary, below we show some other examples without classifying the specific error type.

- *Acervo*:: ‘vacteria’ [sic]; ‘es la consistencia de algunas cosas, p.e. la piña’; ‘afirmar’; ‘contar algo’; ‘hacer muchas cosas’
- *Revelar*: ‘identificar caracteres’; ‘presentar algo a escala’; ‘quitar’; ‘relativo a copiar, algo oculto’; ‘descifrar’.
- *Vario*: ‘del verbo *vasar* [sic]’; ‘*alguna cosa de uso común*’; ‘*mucho*’
- *Vacilo* ‘que no tiene seguridad’; ‘indeciso’

The above presented examples suggest taking some considerations in order to avoid that the requested information cannot be retrieved due to this type of user language fails. We do not expect an information retrieval system tries to “guess” what the user wanted to say from his written query (what he really said).

In the digital version of the Spanish Language Dictionary (Diccionario de la Lengua Española, RAE) there exist a simple mechanism of word ‘approximation’, i.e., if someone writes a word incorrectly (ex. ‘grafica’), it suggest the correct word (‘gráfica’ or ‘graficar’), i.e., it looks for orthographical similar words, which we consider to be a good example of semantic Web.

As conclusions, let us consider the following issues:

- a) The language, any type of language, is a *symbol system*. A system is an ordered set of functional rules of the system items. In this case, we mean as symbols as *those that refer to something, which in general is different of itself and with intentionality*⁵.
- b) The classical analysis⁶ of the verbal sign demonstrate the use of three elements: the *phonetic-acoustic* or significant (and its graphical equivalence, i.e., the letters); the *eidetic-conceptual* or meaning, and the *Referent* or reality, i.e., the target object or what it is referred (it means, it denotes).
- c) The *concept*, the *idea* constitutes the basis of the logic and abstract thought; together with the judgment and argumentation, the complex. The judgment brings together some concepts under certain predication rules (the *sentence* is the *subject predication*), and the argumentation, judgments.

⁵ The term ‘in general’ is due to some signs (verbs) where **the word is the significant thing**. In other words, where *decir* (to say) is *hacer* (to do), like to judge, to promise in contrast with *comer* (to eat), *correr* (to run), etc.

⁶ The **intent** separates symbol or sign or symptom or sign of evidence, not intentional. The same fact can be used as a sign or signal. Let us take for instance the Vatican smoke and that one in a forest, they both ‘indicate’ that there is fire, but the former is deliberate, intentional (it **means** that there are or not a new Pope, white or black, respectively); in the latter case, it **indicates** that there is fire.

⁷ We exclude Saussure and Frege, because we consider them to be non functional. We refer to Ogden and Richards, whom are generally accepted in our environment.

- d) The thought is defined as *rational speech*, and speech is the language as act. It is an update of the potential or human ability of thinking and talking, which implies to learn the reality and not only to extern it (communicative function).
- e) The knowledge is the apprehension of Reality, and this, which includes the principle of identity and space-temporal coordinates, may be defined as *the set of all referents*.
- f) Such apprehension of Reality is performed through the language. It relies under the eidetic-conceptual element, in abstraction and generalization of referents (concrete or abstract, folkloric, geometric, cultural or literary), its reduction to the essential characteristics (different to the accidental ones).
- g) Finally, but no by coincidence, if we follow the Aristotelian logic and categorical grammar, we will find the overlap between grammatical and logical categories, particularly between *ser*(to be)-*sustantivo*(noun); *accidente*(accident)-*adjetivo*(adjective); *acto*(act)-*verbo*(verb).

4 Conclusions

On the basis of the above considerations presented (the thought-language process). This process will depend on the level of system dominion, i.e., to think and link related concepts implies a lexical database and the management of rules in order to link them (grammar). We do not think with images or icons, but with concepts, ideas.

For a simple test, we suggest the reader to think (not to imagine) in a certain reality (abstract or concrete, physical or mental) without appeal to the language (language of any kind) that is no using words.