Concept Maps: Making Learning Meaningful
Proc. of Fourth Int. Conference on Concept Mapping
Viña del Mar, Chile, 2010

# CONCEPT MAPS CORE ELEMENTS CANDIDATES RECOGNITION FROM TEXT

*Juliana H. Kowata, Davidson Cury & Maria Claudia Silva Boeres*

*Universidade Federal do Espírito Santo, Brazil*

*Email: juliana.kowata, dedecury, boeres @gmail.com*

**Abstract.** A concept map is a schematic resource to represent and organize a set of meanings in a propositional structure. In a learning process, the practice of making and remaking concept maps might be considered as an effort to find out concepts and their meanings, giving rise to the knowledge in an explicit way (Novak & Gowin, 1984). Computer aided programs can make the choice to use concept maps easier than before. Over the last few years, many basic functions in concept maps building, such as publishing and sharing, were facilitated by computer aided programs. An increasing interest in applying computational resources to automatically build concept maps from data sources by extracting concepts and linking words has emerged recently. In this paper, we propose an approach focused on the recognition of concept maps' core elements in texts and, in addition, we expose the initial results of the experiment conducted in the Brazilian Portuguese Language.

## 1 Introduction

A concept map is a schematic resource to represent and organize a set of meanings in a propositional structure. In a learning process, the practice of making and remaking concept maps might be considered as an effort to find out concepts and their meanings, giving rise to the knowledge in an explicitly way (Novak & Gowin, 1984).

In general, a concept map springs up from early scratches and much iteration is required to improve it. Hence, it is worth noting that, in some research fields, the process of concept maps building needs to overcome costs and reducing of the time and the efforts in knowledge acquisition activities (Chang et al., 2008; Lee et al., 2009; Tseng et al., 2007), especially, in large knowledge domains highly dependent on experts (Valerio & Leake, 2006). Considering this issues, it is undeniable that computer aided programs work in favor of the practice of thinking using concept maps.

Over the last few years, many basic functions in concept maps building, such as publishing and sharing, were facilitated by computer aided programs. Recently, an increasing concern in applying computational resources to build concept maps automatically from data sources by extracting concepts and linking words have came out. In fact, according to Kowata, Cury & Boeres (2009), 10 of 15 approaches to build concept maps automatically were published in the last three years, most of them (93.34%) concerned with educational (46.67%) and analytical purposes (46.67%). They noticed that 66.67% of the whole approaches used unstructured sources, i.e., natural language texts. With regard to the most applied manipulation methods for unstructured sources, it was observed that only hybrid and linguistic methods produced complete concept maps, representing 33.33%, in counterpart to incomplete maps produced by pure statistical methods.

All exposed information in Kowata, Cury & Boeres (2009) cannot drive us to final conclusions. However, that research outlines a trend, also a challenge, in concept maps building: the growing use of unstructured data sources as primary resource to derive human-understandable concept maps.

With this challenge in mind, we have started a project which aims to define a hybrid approach to handle unstructured data sources in Brazilian Portuguese Language to produce complete concept maps for support learning objects on e-learning environments. In this paper, we propose an approach focused on concept maps core elements recognition in texts. In addition, we expose the initial results of the experiment conducted with a Brazilian Portuguese Language corpus.

The present paper is organized as follows: in Section 2 we go into concept map as a knowledge representation language, highlighting their core elements; in Section 3 we put forward our approach to handle unstructured data source to recognize concept maps core elements candidates; in Section 4 we present an overview of similar approaches; at last, in Section 5 we carry out for pointing up our preliminary findings and our future works.

## 2 Concept Map as Knowledge Representation Language

The concept map language was inspired by Ausubel's Assimilation Learning Theory and Constructivist Epistemology, known as Meaningful Learning Theory. Ausubel stated that the human cognitive structure is an organized and hierarchical space in which the learning process takes place (Novak & Gowin, 1984). During the learning process, new concepts are linked to existent ones by means of propositional statements in a movement called "subsumption" (Cañas et al., 2003). Concept map language exhibits these theoretical assumptions in three major features: a) concepts are organized in a semi-hierarchical way where the more general ones subsume the more specific ones; b) concepts are labeled by a couple of words that defines "a perceived regularity in events or objects, or records of events or objects" (Novak & Gowin, 1984); c) the relationship between two concepts must be labeled in a form so that propositions may be identified.

The two necessary elements of any concept maps are "Concept" and "Relationship". We referrer them as "core elements" to make distinction with resources that was introduced later by some of computer aided tools. Not all concept maps contain additional resources. They act as complimentary resources and comprise format as images, video, web pages, comments or embedded concept maps.

We depict in Figure 1 the concept map language meta-model [1], i.e., it shows the constructs that we can use to build a concept map. We defined that all core elements derive from the same ancestral called "Element". "Element" is a set complete and disjoint, and one instance of element should be or a "Concept" or a "Relationship".

Elements can be enriched by resources and also can be annotated by author's comments. The ontological difference between relationships and concepts are related with their meanings in the real world. While concepts represents "things and events", relationships represents "acts and process". Furthermore, relationships are completely dependents on concepts: in a concept map, the existence of relationships doesn't make sense without "target" and "source" concepts.

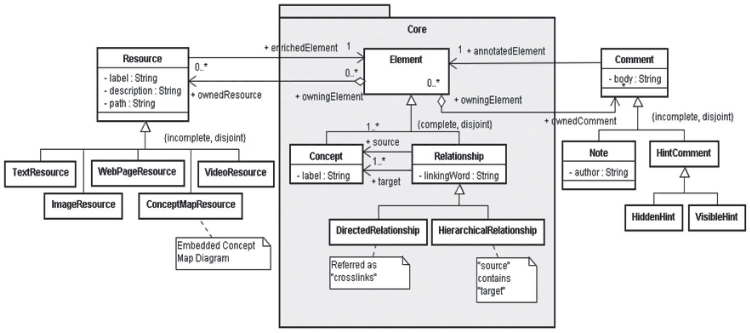


**Figure 1.** Concept Map Language Meta-model

## 3 Concept Maps Core Elements Candidates Recognition

This research was conceived with two fundamental assumptions: (i) try to put forward an approach that could work with unstructured data source and (ii) the resulted concept maps should be understandable to humans. The first assumption related to "work with unstructured data source" assumes that the proposed approach have to face the complexity of the natural language, while the second one, about "produce understandable output to humans", means that

[1] A meta-model is composed by language and rules that define the process and elements of the model.

the process of building concept map from a text should offer minimum semantic losses and preserve the main idea of the source.

For achieving results that could be adherent to the defined assumptions, we try to propose a whole approach with several minor solutions, each one concerned with a single problem. Actually, we design a pipeline with a set of coordinated tasks in which each task depends on the output of a previous task and produce an output used in the next task, as depicted in Figure 2.
related to "work with unstructured data source" assumes that the proposed approach have to face the complexity of the natural language, while the second one, being comprehensible, means that the process of building concept map from a text should offer minimum semantic losses and preserve the main idea of the source.

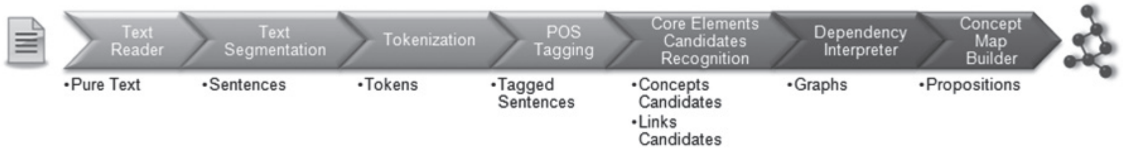


**Figure 2.** Pipeline to build Concept Map from text

In "Learning How to Learn", Novak & Gowin (1984) stated that the choice of words is very important to label concepts and relationships. Keeping in mind that sentences are a list of words arranged under syntactical rules, we focused on finding out the more suitable subset of these words to name the core elements of concept maps.

We defined a metaphor "disassemble to reassemble" to guide the search for concept and relationships from sentences in natural language. Step by step, we gradually broke up a text until its structure became visible, emphasizing the "disassembling" stage: at first by removing format tags in order to find out a pure text, then by splitting the text in sentences, and at last, by identifying independent tokens. In turn, each token was identified with its Part-of-Speech tag[2] (POS tag). From this point on, another bias would take place, highlighting the "reassembling" stage. Using a set of linguistic patterns based on POS tag, specific linguist chunks[3] were recognized in an activity called "Core Elements Candidate Recognition". In next stages, candidates were mapped into graph[4] nodes. Just over another transformation from graph to concept map, we can find out "concepts" and "relationships", as depicted in Figure 3.

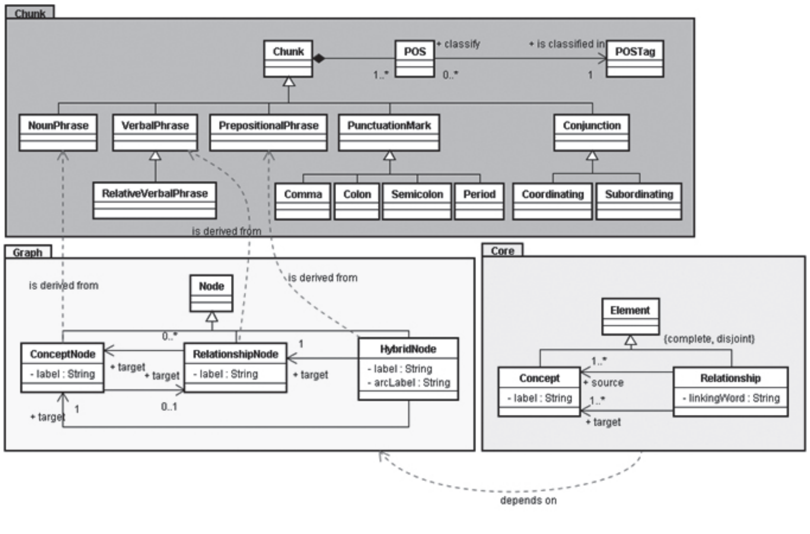


**Figure 3.** Mapping Chunks, Graphs and Core Elements

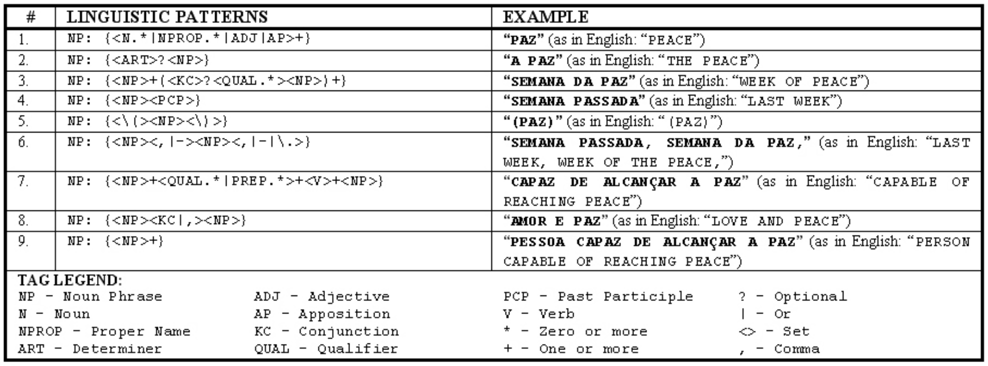
[2] Part-of-Speech tag (POS tag) is a label that identifies a "word class", for instance, nouns, verbs, adjectives, adverbs, prepositions, etc.
[3] Linguistic chunk is a sequence of words that make sense when disposed together. Among these words there is only one that is indispensable, often, a noun or verb.
[4] Graph is a mathematical structure composed by nodes and arcs

### 3.1 Core Elements Candidates

As in English Language, in Brazilian Portuguese it is also possible to identify sequences of words surrounding at least one noun (Noun Phrase), verb (Verbal Phrase), or preceded by a preposition (Prepositional Phrase). These sequences compound what we call "linguistic chunks".

According to Novak & Gowin (1984) it is essential to isolate concept from linking words (relationships). Both of them are important as language units, but they play different roles in the transmission of meaning. While concept is labeled by words that represent things and events, relationships are a way to link two concepts in a propositional form. Novak & Gowin (1984) did not define any correlation between concept maps elements and linguistic chunks. However, according to their directions to build a concept map, we could consider Noun Phrase (NP), Verbal Phrase (VP) and Prepositional Phrase (PP) as primary candidates to be mapped as Concept Maps Core Elements. Noun Phrase (NP) is the most important candidate to be a concept; Verbal Phrase (VP) is aspirant to be a relationship, and Prepositional Phrase (PP) could be both a concept, taking the essence of the Noun Phrase, and a relationship, if we consider the preposition that appears before the Noun Phrase.

As a general rule, sentence chunking is an important task for recognizing the core elements candidates. Chunks are created by a set of linguistic patterns, each of them described by regular expressions, in Brazilian Portuguese language. An example of linguistic patterns for NP recognition in Portuguese Language can be seen in Table 1.



| # | LINGUISTIC PATTERNS | EXAMPLE |
|---|---|---|
| 1. | NP: {<N.*\|NPROP.*\|ADJ\|AP>+} | **"PAZ"** (as in English: "PEACE") |
| 2. | NP: {<ART>?<NP>} | **"A PAZ"** (as in English: "THE PEACE") |
| 3. | NP: {<NP>+(<KC>?<QUAL.*><NP>)+} | **"SEMANA DA PAZ"** (as in English: "WEEK OF PEACE") |
| 4. | NP: {<NP><PCP>} | **"SEMANA PASSADA"** (as in English: "LAST WEEK") |
| 5. | NP: {<\(><NP><\)>} | **"(PAZ)"** (as in English: "(PAZ)") |
| 6. | NP: {<NP><,\|-><NP><,\|-\|\.>} | **"SEMANA PASSADA, SEMANA DA PAZ,"** (as in English: "LAST WEEK, WEEK OF THE PEACE,") |
| 7. | NP: {<NP>+<QUAL.*\|PREP.*>+<V>+<NP>} | **"CAPAZ DE ALCANÇAR A PAZ"** (as in English: "CAPABLE OF REACHING PEACE") |
| 8. | NP: {<NP><KC\|,><NP>} | **"AMOR E PAZ"** (as in English: "LOVE AND PEACE") |
| 9. | NP: {<NP>+} | **"PESSOA CAPAZ DE ALCANÇAR A PAZ"** (as in English: "PERSON CAPABLE OF REACHING PEACE") |

**TAG LEGEND:**

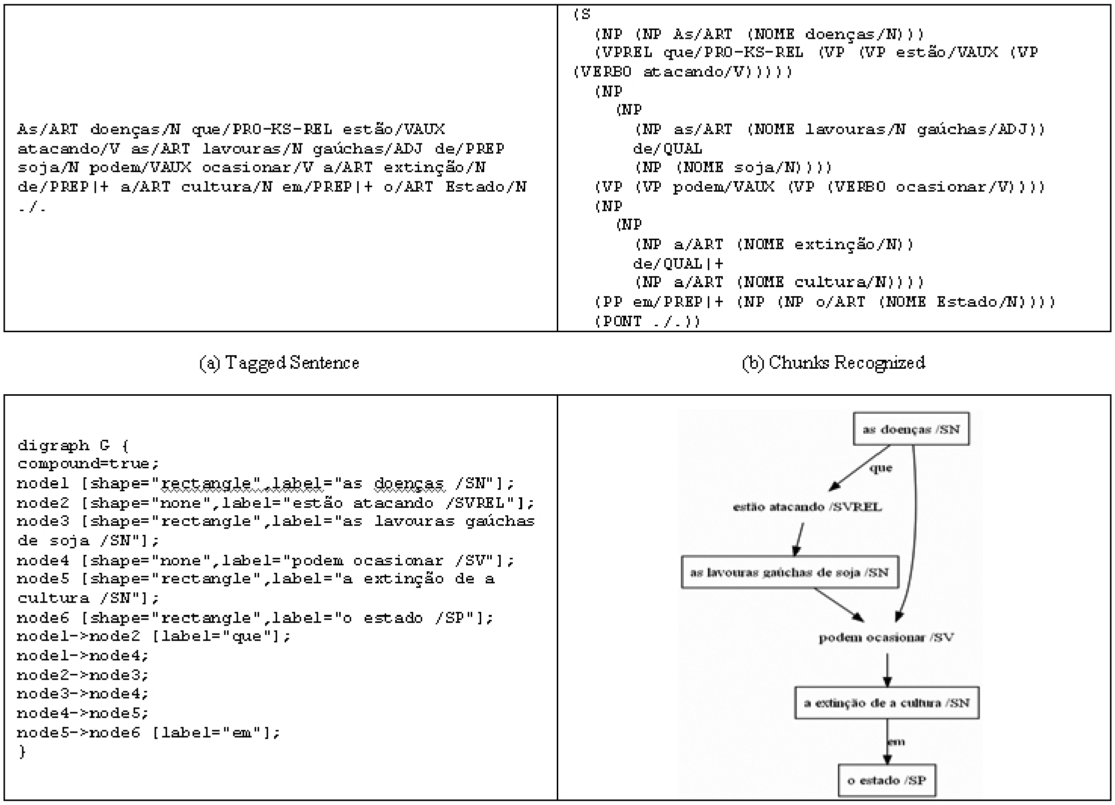
| | | | |
|---|---|---|---|
| NP - Noun Phrase | ADJ - Adjective | PCP - Past Participle | ? - Optional |
| N - Noun | AP - Apposition | V - Verb | \| - Or |
| NPROP - Proper Name | KC - Conjunction | * - Zero or more | <> - Set |
| ART - Determiner | QUAL - Qualifier | + - One or more | , - Comma |

**Table 1.** Linguist Patterns for NP Recognition in Brazilian Portuguese Language

In the Core Elements Candidates Recognition task, a program takes a tagged sentence (Figure 4a) and searches for all structures that match with the defined patterns. This is performed recursively until each production rule has been processed. At the end, specific chunks are identified (Figure 4b), such as: Noun Phrases, Verbal Phrases, Prepositional Phrases, Punctuation Marks, and Conjunctions etc.

Each chunk has a particular behavior when it is put together with another chunk, for example: VP always wants to attach an NP above it, but, it usually ignores a PP that appears in the same condition. These behaviors were codified into a program that, with some rules, identifies dependencies between the chunks. When a binary dependency is identified, each chunk instance is mapped into a graph node instance and the relation between them is explicitly represented by a labeled arc. Each NP is mapped into an instance of ConceptNode, VP into RelationshipNode, and PP into HybridNode, as depicted by Graph Package in Figure 3. This way, a graph is built to be an intermediate structure between a text and a concept map.

Graphs were chosen as an intermediate structure, because their form with nodes and arcs is very similar to a concept map. Besides that, they don't have constraints to deal with binary relations. From this point of view, graphs are more flexible than concept maps, considering that in the latter there is the requirement to produce propositions. As we can observe, in graphs, a relationship could also be represented as a node in the same way as a concept. Therefore, to transform the graph nodes into concept maps elements is an indispensable stage to build a concept map.



```
As/ART doenças/N que/PRO-KS-REL estão/VAUX
atacando/V as/ART lavouras/N gaúchas/ADJ de/PREP
soja/N podem/VAUX ocasionar/V a/ART extinção/N
de/PREP|+ a/ART cultura/N em/PREP|+ o/ART Estado/N
./.
```

(a) Tagged Sentence

```
(S
  (NP (NP As/ART (NOME doenças/N)))
  (VPREL que/PRO-KS-REL (VP (VP estão/VAUX (VP
(VERBO atacando/V)))))
  (NP
    (NP
      (NP as/ART (NOME lavouras/N gaúchas/ADJ))
      de/QUAL
      (NP (NOME soja/N))))
  (VP (VP podem/VAUX (VP (VERBO ocasionar/V))))
  (NP
    (NP
      (NP a/ART (NOME extinção/N))
      de/QUAL|+
      (NP a/ART (NOME cultura/N))))
  (PP em/PREP|+ (NP (NP o/ART (NOME Estado/N))))
  (PONT ./.))
```

(b) Chunks Recognized

```
digraph G {
compound=true;
node1 [shape="rectangle",label="as doenças /SN"];
node2 [shape="none",label="estão atacando /SVREL"];
node3 [shape="rectangle",label="as lavouras gaúchas
de soja /SN"];
node4 [shape="none",label="podem ocasionar /SV"];
node5 [shape="rectangle",label="a extinção de a
cultura /SN"];
node6 [shape="rectangle",label="o estado /SP"];
node1->node2 [label="que"];
node1->node4;
node2->node3;
node3->node4;
node4->node5;
node5->node6 [label="em"];
}
```

(c) Graph in dot Format

(d) Graph

**Figure 4.** Pipeline's outputs

## 4 Case study

We selected MAC-MORPHO, which uses Brazilian Portuguese Language, from the Lácio-Web Project (Aluisio et al, 2004), to test the approach. This linguistic resource comprises approximately 1.2 million words, which was manually-validated for morpho-syntactical tags, distributed in 51,397 annotated sentences.

From the body of MAC-MORPHO, we selected 2% of texts, splitting each of them in sentences. As a result, we had 927 sentences processed, 927 graphs produced, 1710 VP, 2059 NP, and 1505 PP identified. These numbers correspond to an approximately 1.84 VP, 2.22 NP, 1.62 PP, and an average of 21.54 tokens per sentence. In general, each sentence, in average, was reduced to 4 main chunks (NP and VP) and 1 complement (PP).

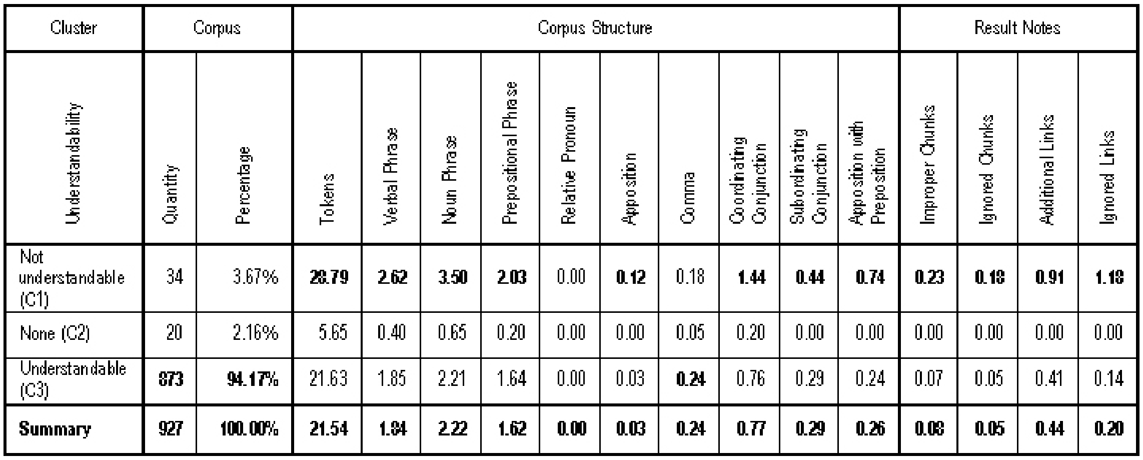
The graph building program was codified and compiled using Python 2.6.5 (Python, 2010), using Natural Language Processing Toolkit (Bird, Klein & Loper, 2009) and GvGen (GvGen, 2007) – a Python module used to generate Graphviz-compatible dot files. To visualize the graphs, we used Graphviz (Graphviz, 2010).

### *4.1 Preliminary Analysis*

The 927 results were classified in three major clusters, according to the graph comprehensibility. The clusters are: C1 – Not Understandable, C2 – None, C3 – Understandable.

The clustering was conducted by human experts that have compared the original source with the graph result. Using a simple checklist, they classified each of the 927 cases in only one cluster. In addition, it was demanded that notes were gathered, such as the number of improper chunks, of ignored chunks, of additional links, and ignored links.

A summary of the work can be seen in Table 2 where we can observe that 3.67% of the cases were classified as not readable while 94.17% were considered readable. In 2.16% the program could not produce a graph output.



| Cluster | Corpus | | Corpus Structure | | | | | | | | | | Result Notes | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Understandability | Quantity | Percentage | Tokens | Verbal Phrase | Noun Phrase | Prepositional Phrase | Relative Pronoun | Apposition | Comma | Coordinating Conjunction | Subordinating Conjunction | Apposition with Preposition | Improper Chunks | Ignored Chunks | Additional Links | Ignored Links |
| Not understandable (C1) | 34 | 3.67% | **28.79** | **2.62** | **3.50** | **2.03** | 0.00 | **0.12** | 0.18 | **1.44** | **0.44** | **0.74** | **0.23** | **0.18** | **0.91** | **1.18** |
| None (C2) | 20 | 2.16% | 5.65 | 0.40 | 0.65 | 0.20 | 0.00 | 0.00 | 0.05 | 0.20 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Understandable (C3) | **873** | **94.17%** | 21.63 | 1.85 | 2.21 | 1.64 | 0.00 | 0.03 | **0.24** | 0.76 | 0.29 | 0.24 | 0.07 | 0.05 | 0.41 | 0.14 |
| **Summary** | **927** | **100.00%** | **21.54** | **1.84** | **2.22** | **1.62** | **0.00** | **0.03** | **0.24** | **0.77** | **0.29** | **0.26** | **0.08** | **0.05** | **0.44** | **0.20** |

**Table 2.** Summary of Preliminary Analysis

In the C1 cluster, we found out that the body of the input has high occurrences of tokens, chunks and misunderstood chunks in comparison to C2 and C3 cluster. We could infer that this, in some way, influenced the semantic quality of the results.

The C2 cluster represents cases where it was not possible to recognize links between elements of chunks. In fact, the links depend on the existence of verbal phrases. This cluster has low average of tokens, especially of verbal phrases.

The C3 cluster contains understandable structures in comparison with the source sentences, representing 94.17% of the whole sample.

The preliminary results show that the approach was successful to handle the major part of the body [of language] in a satisfactory way. Considering the results, we could observe that the approach worked successfully with incomplete sentences, subordinated and coordinated forms, as well with topicalized forms. However, lots of work remains to be done to improve the program performance. It still needs to evolve and to be more robust in dealing with the anaphora phenomena, apposition and interrogative as well as imperative sentences.

## 5 Similar Approaches

The use of linguistic methods to build a concept map from text is not a new approach (Kowata, Cury & Boeres, 2009). In fact, Valerio & Leake (2006) also used Noun Phrases and Verbal Phrases to extract concepts and relationships. In another way, Pérez & Vieira (2004; 2005) proposed to handle syntactic chunks as subject, predicate and objects and mapped them to core elements by using a set of rules. Alves, Pereira & Cardoso (2001) used linguistic methods associated to machine learning methods to create binary predicates based on external lexical resource. Lau et al. (2008) used forum and chat texts and extracted concept maps from them by applying linguistic patterns and statistical methods to compute the relevance of each concept. Richardson & Fox (2007) and Richardson, Srinivasan & Fox (2008) extracted propositions from theses and dissertations by using a tool to handle texts and extract information of them. Zouaq & Nkambou (2008, 2009) extracted key concepts from natural language text using machine learning and linguistic methods.

The most relevant difference between the mentioned approaches and the approach presented in this paper are:

a) we aim to produce one concept map from one text, in counterpart of Valerio & Leake (2006) and Zouaq & Nkambou (2008, 2009) that retrieves information of more than only one text;

b) we propose the use of raw text as input of concept maps building approach, differently from Pérez & Vieira (2004, 2005) that require pre-tagged texts;

c) we don't ignore stop words – actually we consider that it is very important to define the dependency between concepts and relationships –, such as prepositions and punctuations, as we can verify in Lau et al. (2008);

d) we aim to produce comprehensible concept maps and a concept map is our primary goal, differently from Zouaq & Nkambou (2008, 2009) that uses concept maps as intermediate language to build ontologies;

e) we do not restrict the set of linking words, which is in opposition to Alves, Pereira & Cardoso (2001).

To sum up, it is important to highlight that we propose an approach to build concept maps from text, describing each task in detail and offering a single tool that supports all the tasks, in a clear way to the end user.

## 6 Future Works

We have noticed that the heuristic used to build dependencies between graphs nodes could present better performance. Thus, it is necessary to improve it or define others that could deal better with ambiguity of prepositional phrases, anaphora phenomena, apposition and interrogative and imperative sentences.

Future works also include the evaluation of taggers for Brazilian Portuguese Language in order to select one more suitable to the research needs and the development of a prototypical tool to build concept maps automatically from the inputted text.

## References

Aluisio, S., Pinheiro, G. M., Manfrim, A. M. P., Oliveira, L. H. M. de., Genoves Jr., L. C. , Tagnin, S. E. O. (2004). The Lácio-Web: Corpora and Tools to advance Brazilian Portuguese Language Investigations and Computational Linguistic Tools. Paper presented at the 4th International Conference on Language Resources and Evaluation (LREC 2004).

Alves, A. O., Pereira, F. C., Cardoso, A. (2001). Automatic Reading and Learning from Text. In: Proceedings of the International Symposium on Artificial Intelligence (ISAI'2001), (pp. 302-310). Fort Panhala (Kolhapur), India.

Bird, S., Klein, E., Loper, E. (2009). Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit. CA: O'Reilly Media.

Cañas, A. J., Coffey, J. W., Carnot, M. J., Feltovich, P., Hoffman, R. R., Feltovich, J., et al. (2003a). Summary of Literature Pertaining to the Use of Concept Mapping Techniques and Technologies for Education and Performance Support. Acesso em 6 de junho de 2009, disponível em http://www.ihmc.us/users/acanas/Publications/ConceptMapLitReview/IHMC%20Literature%20Review%20on%20Concept%20Mapping.pdf

Chang, T.-H., Tam, H.-P., Lee, C.-H., Sung, Y.-T. (2008). Automatic Concept Map Constructing using top-specific training corpus. Paper presented at the Asia-Pacific Educational Research Association Board Meeting (APERA'2008).

Graphviz 2.26.3 (2010). Graphviz Project Website. Access date: 2010-03-22, download available at http://www.graphviz.org/Download_windows.php.

GvGen 0.9. (2007). GvGen Project Website. Access date: 2010-03-22, download available at http://software.inl.fr/trac/wiki/GvGen.

Kowata, J. H., Cury, D., Boeres, M. C. S. (2009). Caracterização das Abordagens para Construção de Mapas Conceituais. Paper presented at the XX Brazilian Symposium on Computer in Education (SBIE 2009).

Lee, C.-H., Lee, G.-G., Leu, Y. (2009). Application of automatically constructed concept map of learning to conceptual diagnosis of e-learning. In: Expert Systems with Applications , 36 (2), 1675-1684.

Novak, J. D., & Gowin, D. B. (1984). Learning How to Learn. New York: Cambridge University Press.

Pérez, C. C., Vieira, R. (2004). Aquisição de Conhecimento a partir de Textos para Construção de Mapas Conceituais. In: II Workshop de Teses e Dissertações em Inteligência Artificial (WTDIA 2004). São Luís, MA.

Pérez, C. C., Vieira, R. (2005). Mapas Conceituais: geração e avaliação. In: Anais do III Workshop em Tecnologia da Informação e da Linguagem Humana (TIL'2005), (pp. 2158-2167). São Leopoldo, RS.

Python 2.6.5. (2010). Python Programming Language Official Website. Access date: 2010-03-22, download available at http://www.python.org/download/releases/2.6.5/.

Richardson, R., Fox, E. A. (2007). Using Concept Maps in NDLTD as a Cross-Language. In: 10th International Symposium on Electronic Theses and Dissertations (ETD 2007). Uppsala, Sweden.

Richardson, W. R., Srinivasan, V., Fox, E. A. (2008). Knowledge discovery in digital libraries of electronic theses and dissertations: an NDLTD case study. In: International Journal on Digital Libraries , 9 (2), 163-171.

Tseng, S.-S., Sue, P.-C., Su, J.-M., Weng, J.-F., Tsai, W.-N. (2007). A new approach for constructing the concept map. In: Computers & Education , 49 (3), 691-707.

Tseng, S.-S., Sue, P.-C., Su, J.-M., Weng, J.-F., Tsai, W.-N. (2007). A new approach for constructing the concept map. In: Computers & Education , 49 (3), 691-707.

Valerio, A., Leake, D. (2006). Jump-Starting Concept Map Construction with Knowledge Extracted from Documents. In: A. J. Cañas, J. D. Novak, F. M. González (Ed.), Paper presented at the Second International Conference on Concept Mapping (CMC'06).

Zouaq, A., Nkambou, R. (2008). Building Domain Ontologies from Text for Educational Purposes. In: IEEE Transactions on Learning Technologies , Volume 1 (1), p. 49-62.

Zouaq, A., Nkambou, R. (2009). Evaluating the Generation of Domain Ontologies in the Knowledge Puzzle Project. In: IEEE Transactions on Knowledge and Data Engineering (10.1109/TKDE.2009.25).