Tira-Teimas: after Shallow Parsing

Luísa Coheur*,+, Nuno J. Mamede*

*L2F INESC-ID/IST - Spoken Languages Systems Laboratory
+GRIL/Université Blaise-Pascal
Rua Alves Redol nº 9, 1000 - 029
{luisa.coheur}{numo.mamede}@l2f.inesc-id.pt

Abstract

In this paper we present Tira-Teimas, which is a program written in XSLT, that checks if a shallow parsed text verifies a set of properties from the 5P paradigm. We show how to code exigency properties in XSLT and we present an example of a model disrespecting a property.

1. Introduction

Tira-Teimas is a program that verifies if a shallow parsed text satisfies a set of properties from the 5P paradigm (Bès, 99; Bès & Hagège, 2001; Hagège, 2000). These properties are used to describe the syntax of a natural language.

Tira-Teimas checks the following properties (concerning a family of models (phrases), labelled M):

- Uniqueness: identifies the elements that cannot occur more than once in a model labelled M;
- Exigency: allows to declare that a occurs in a model labelled M only if b also occurs in it;
- Exclusion: permits to declare that a excludes b in a model labelled M;
- Linearity: declares the linearity relations between the elements occurring in a model labelled M.

These properties can be seen as a repository of linguistic information that can be used according to our needs (Bès & Hagège, 2002). Having nominal phrases extraction as a goal, Hagège developed a shallow parser prototype, AF, that uses information from the 5P properties (Hagège, 2000; Bès & al., 1999). Therefore, 5P properties for nominal models (Hagège, 2000) enriched the information structures used by AF. Nevertheless, these information structures are less expressive than the 5P properties. Therefore, it is not sure that the models identified by AF verify the whole set of 5P properties. As so, **Tira-Teimas** was developed in order to verify if each model identified by AF satisfies (or not) the 5P properties.

2. Tira-Teimas

Tira-Teimas is written in XSLT (W3C-XSL). The following example shows how to code exigency properties in a format that can be easily mapped into XSLT (a similar approach can be applied to other 5P properties).

Exigency properties have the following general syntax:

$$E_i: \{a_1, ..., a_n\} \Rightarrow_M \{b_1, ..., b_m\} \mid ... \mid \{c_1, ..., c_k\}$$

meaning that if the symbols $a_1, ..., a_n$ occur in a model labelled M, then

$$\{b_1, ..., b_m\} \text{ or ... or } \{c_1, ..., c_k\}$$

must also occur in that model.[1]

Given this, **Tira-Teimas** works as follows: suppose that a model labelled M is detected by the shallow parser. Then **Tira-Teimas** checks if it verifies E_i by counting the number of occurrences of every a_j, b_k and c_m in the model. That is, being count(x, X) a function returning the number of occurrences of x in (the model) X, if

$$\text{count}(a_1, M) \mathrel{!=} 0 \text{ and ... and count}(a_n, M) \mathrel{!=} 0$$

and

$$[(\text{count}(b_1, M) = 0 \text{ or ... or count}(b_m, M) = 0)$$

and ... and

$$(\text{count}(c_1, M) = 0 \text{ or ... or count}(c_k, M) = 0)]$$

then M does not satisfies E_i.

When a model does not satisfies a property, it is marked with the identification of that property.

As a predicate count is available in XSLT, mapping the previous formulas in XSLT is a trivial task. On the contrary, writing 5P properties directly in XSLT is not an easy task, as 5P properties in XSLT take a very

[1] These elements or others subsumed by them .

unfriendly look. In order to solve this situation, 5P properties are written in XML (W3C-XML). Then an extra program **TTT** (Tira-Teimas Translator), maps these properties into XSLT. **TTT** is also written in XSLT.

3. Results

SuSAna (Batista & Mamede, 2002) is, in rough terms, a new implementation of AF, that we used to collect a set of shallow parsed corpus. Experiments with **Tira-Teimas** were made over these corpora. As expected a few (not many) models disrespected 5P properties. The following example describes a situation where a model does not verifies a property.

Consider exigency property E_{15} from (Hagège, 2000), over nuclear nominal models (labelled m-nn)[1]:

E_{15} adj_s $\Rightarrow_{nn}$ det | cada | qualquer | certo1 | algum | nenhum | tal | outro | tanto

The linguistic information that SuSAna uses, accepts that inside an m-nn, *muito* (labelled q3_s) can be followed by an adj1_s (consider for example *Ele comeu muito belo peixe*. *Tanto que ficou doente*.[2]).

As so, the following model was captured by SuSAna:

$(\textit{muito}_{q3\_s}\ \textit{cansado}_{adj1\_s})_{nn}$[3]

Tira-Teimas ran over the same corpus and detected an inconsitency between E_{15} and that syntactic model. In fact, it does not respect E_{15}, because according to E_{15} adj1_s (subsumed by adj_s) requires one of the elements on the right side of E_{15}, and q3_s is not one of them.

4. Conclusions and future work

Although the original motivation for **Tira-Teimas** was to check 5P properties, **Tira-Teimas** is not bounded to this application.

In fact, it can also be used to find differences between what is syntactically correct - supposing that a set of 5P properties describe it - and what is currently practised.

In addition, **Tira-Teimas** could be applied to detect differences between the Portuguese from Portugal and Portuguese from Brazil. For example, the 5P properties from (Hagège, 2000) describing Portuguese (from Portugal) nominal phrases could be applied to a Brazilian shallow parsed text.

Finally, **Tira-Teimas** can be easily extended to other properties.

5. Acknowledgments

Paper supported by FCT (Fundação para a Ciência e Tecnologia).

6. References

Batista, F., Mamede N. 2002. SuSAna: Módulo multifuncional da análise sintáctica de superfície. In J. Gonzalo, A. Penas, and A. Ferràndez (eds.), *Proc. Multilingual Information Access and Natural Language Processing Workshop*, pages 29-37, Sevilla, Sapain, November 2002. IBERAMIA 2002.

Bès, G. G., Hagège, C., 2001. Properties in 5P (soon in the GRIL web page). Technical Report, GRIL, Clermont-Ferrand, France, November, 2001.

Bès, G. G., 1999. La phrase verbal noyau en français. In *Recherches sur le français parlé*.15: 273-358. Université de Provence, France, 1999.

Bès, G. G., Hagège, C., Coheur L., 2001. Des propriétés linguistiques à l'analyse d'une langue. In *VEXTAL*. Venice, Italy, November, 1999.

Hagège, C., 2000. *Analyse Syntatic Automatique du Portugais*. Ph.D. Thesis, Université Blaise-Pascal, Clermont-Ferrand, France, 2000.

Hagège, C., Bès, G. G., 2002. Enconding and reusing linguistic information expressed by linguistic properties. In *Proceedings of COLING*'2002. Taipei, 2002.

World Wide Web Consortium (W3C). *Extensible Markup Language (XML)*. See: www.w3.org/XML.

World Wide Web Consortium (W3C). *The Extensible Stylesheet Language (XSL)*. See: www.w3.org/Style/XSL.

[1] adj_s is the label of a category subsuming adj1_s, adj2_s and adj3_s - adjectives of type 1, 2 and 3, respectively. Det stands for determiners, and cada, qualquer, certo1, algum, nenhum, tal, outro, tanto are very particular category labels for the words *cada*, *qualquer*, *certo1*, *algum*, *nenhum*, *tal*, *outro* and *tanto*, respectively.

[2] *He ate lots of nice fish. And he became sick.*

[3] *Muito cansado* means *very tired.*