Some Families of Compound Temporal Adverbs in Portuguese

Jorge BAPTISTA
Universidade do Algarve –FCHS / CAUTL – LabEL, IST
Campus de Gambelas
Faro, Portugal, P–8000–117
jbaptis@ualg.pt

Abstract

This paper deals with some families of compound temporal adverbs in Portuguese, formed around the time-related noun *ano* (year). It focuses on several problems arising during the description of formal variation of these expressions by means of finite-state methods in view of their automatic processing.

Introduction

Temporal adverbs constitute an important part of the meaning units of texts, especially in informative-narrative discourses such as newspaper news. Identifying the lexical units of a text, both simple and compound, is the first step towards its automatic processing. Usually, this can be done by means of electronic dictionaries.

Significant lexical coverage of simple word adverbs (i.e., formed by a single word), can already be found in some dictionaries, in spite of difficulties arising from some morphologically productive classes, e.g. *-ly* adverbs: *suddenly*, Molinier and Levrier (2000). However, compound adverbs, i.e., adverbs formed by two or more simple words such as: *as soon as possible*, *by no means*, *from time to time*, *from now on*, are more difficult to list in full, since many of them do not appear in reference dictionaries – see M. Gross (1982, 1986) for a formal definition and a comprehensive description.

Usually, these adverbs are syntactically frozen and their meaning can not be calculated from the meaning of their component words. Due to the idiosyncrasy of these lexical combinations, there seems to be no other way but to build electronic dictionaries of frozen adverbs in order to recognize and adequately tag them in texts (Ranchhod *et al.* 1999).

Many time-related compound adverbs are formed around *Ntmp*: *second*, *minute*, *hour*, *day*, *week*, *year*, *century*, *morning*, *afternoon*, *evening*, *eve*, *moment*, *instant*, *time*, an so on. Often, an entire family of adverbs can be found formed around a time-related noun (*Ntmp*). For instance, the following are some of the adverbs formed around the *Ntmp* ***moment***: *for a moment*, *for the moment*, *in a moment*, *moments latter*, *not a moment too soon*, *on the spur of the moment*, etc.

In some cases, adverbial phrases with *Ntmp* form rather complex and often productive combinations; often, these linguistic expressions tend to be semantically transparent but syntactically constraint. This is the case of date expressions: *on January 1st, 2003*, *on the eve of April 11th, 2003* – Maurel (1990, 1992), Martinez-Barco *et al.* (2002); or combinations of ‘dates’, ‘hour’ and the different parts of the day: *yesterday morning*, *late in the evening*, *at two o’clock in the afternoon*, – Baptista (1999); Baptista and Català Guitart (2002).

Due to the sheer size of the combinations involved (several thousand different expressions), it would be impractical to list them all in a dictionary. Considering (a) the fact that they obey a rather constrained composition pattern; (b) their modularity and combinability; and (c) the relative independence from the sentences in which they appear – at least as far as their formal recognition is concerned; it seems to be more efficient and adequate to represent these linguistic expressions by means of finite-state methods (M. Gross 1997).

This paper reports on on-going research on complex time-related adverbial phrases in Portuguese. It follows the methodology of describing them by systematically exploring the combinations in which a *Ntmp* or a *Ntmp* family appear. This paper will deal with the *Ntmp ano* (year). It will focus on several problems regarding the formal complexity of these linguistic expressions in view of their automatic processing. These aspects are also to be found in similar expressions with other *Ntmp*.

1 Methods

As a source for raw data, a corpus of Portuguese journalistic texts – the *CETEMPúblico* corpus [1] – was explored with the *INTEX* (Silberztein 1993, 2000) linguistic development platform [2] and the linguistic resources developed for Portuguese by the LabEL team (Eleutério *et al.* 1995; Ranchhod *et al*. 1999).

INTEX was also used to build the finite-state transducers (FST) that describe the compound adverbs studied here. The FST are used to identify and tag the adverbs in texts. In the graphical representation of FST adopted in this system (Fig. 1): states are left implicit and the input symbols (words) are inside the nodes, whereas the output symbols (linguistic information) are below the nodes; auxiliary graphs appear in grey boxes [3].

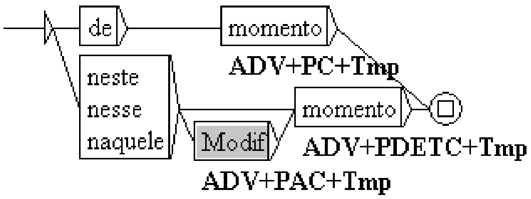

Fig.1. Example of the graphical representation of a finite-state transducer in *INTEX*.

Using M. Gross's (1986: 11) notion of *generalized adverb*, adverbial phrases studied here follow the basic structure (A):

(A) ***Prep Det N Modif***

i.e. a preposition (*Prep*), a determiner (*Det*), the noun (*N*) *ano* and a modifier (*Modif*) [4]. As a guiding method for systematically exploring the combinations in which the *Ntmp ano* appears [5], we begin with the simplest combinations and proceed to the more complex ones.

In order to take into account number agreement, expressions with singular form *ano* were sometimes described separately from those with the plural form *anos*.

1.1. Compond *Ntmp* with *ano*

The noun *ano* is found in many compound nouns that may function as a *Ntmp* and may alternate with the simple noun: *ano civil*, *ano económico*, *ano fiscal*, *ano lectivo*.

1.2 Prepositions

FST libraries were organized by the initial *Prep* [6]. For these, we have only considered prepositions that, in combination with *ano*, express **basic** temporal localization: *em* (in), **duration**: *durante*, *ao longo de* (during), *por* (by), **beginning**: *de*, *desde*, *a partir de* (from); **end**: *até* (until); and approximate indications: *perto de* (near) [7].

1.3 Determiners and modifiers

For each *Prep ano* combination a list of possible determiners and modifiers was established, as well as combinatorial constrains between them. The next FST (Fig. 2) illustrate some of the

[1] *http://cgi.portugues.mct.pt/cetempublico*. This is a corpus obtained from the daily newspaper Público. We only used the first fragment of this corpus, which constitutes a text of 58,7 Mb and contains 9,6 million words. In this fragment, the noun *ano* appears 29,118 times, including all inflected forms, plural *anos* and diminutives *aninho* and *anito*. For clarity, these diminutives were left out in the FST presented in this paper.

[2] *http://www.bestweb.net/~intex*.

[3] This FST identifies compound adverbs such as *de momento* (*at the moment*) and attributes the tag **ADV** (= adverb), its formal class (**PC**, **PDETC**, **PCA**) following the classification principles of M. Gross (1982, 1986), and the feature **Tmp** (= time). Auxiliary graph *Modif* contains a list of modifiers that appear in expressions such as *nesse* (E + *preciso*) *momento* (*at that precise moment*).

[4] Some adverbs do not present the initial preposition: *este ano* (this year); others do not have both determiner and modifier, *por ano* (per year).

[5] Combinations involving year values but without the *Ntmp ano*, e.g. *<Isso aconteceu> em 2003* (*<That happened> in 2003*), were not taken in consideration, since they were often ambiguous.

[6] In some cases, other words were taken as the organizing motif, as in the adverbs formed with the impersonal verb *haver* (there be); see below.

[7] The English translation of words and examples is aproximate, often literal, and is only meant to show the syntactic phenomena ; its acceptability is irrelevant for the purpose of this paper.

more common *Det – Modif* combinations in adverbs beginning with Prep *em* :

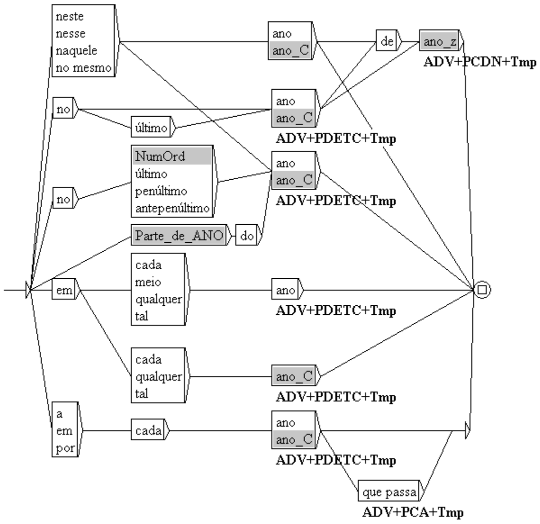

Fig. 2. *EmDetAno.grf*

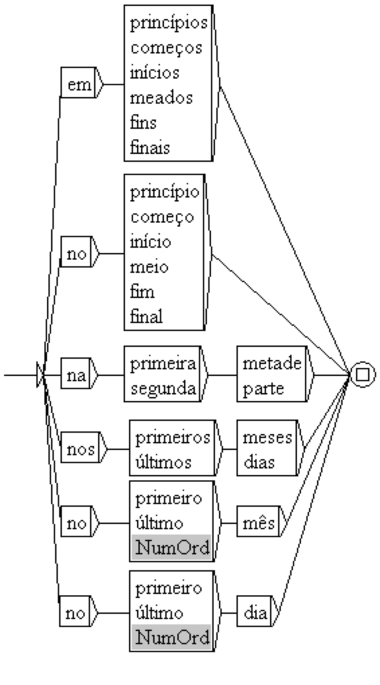

Fig. 3. *Parte_de_ANO.grf*

In this FST, subgraph **ano_C** contains compound *Ntmp* formed with the noun *ano*. Subgraph **ano_z** represent year numbers (*2003*) and several ways of expressing year ranges (*2000-2003*, *2002-2003*, *2002/03*, *2002 a 2003*). **NumOrd** describes ordinal numerals, both simple and compound: *primeiro* (first), *segundo* (second), *vigésimo terceiro* (23rd). A set of determiners that refer to parts of the year was described in subgraph **Parte_de_ANO** (Fig. 3).

In some cases, an effort was made to describe even more complex determiner-modifier combinations. This was done, for instance, with the family of comparative expressions:

<Tal como aconteceu> em igual período do ano passado
(*<As it happened> in the same period of the previous year*)

where certain relative clauses can frequently occupy the syntactic position of *Modif:*

<Tal como aconteceu> em igual período do ano que (E + *agora*) (*termina* + *acaba* + *finda* + *começa* + *se inicia*)
(*<As it happened> in the same period of the year that is now beguining / ending*)

The most frequent of these relative clauses were represented in the FST (Fig. 4, first line):

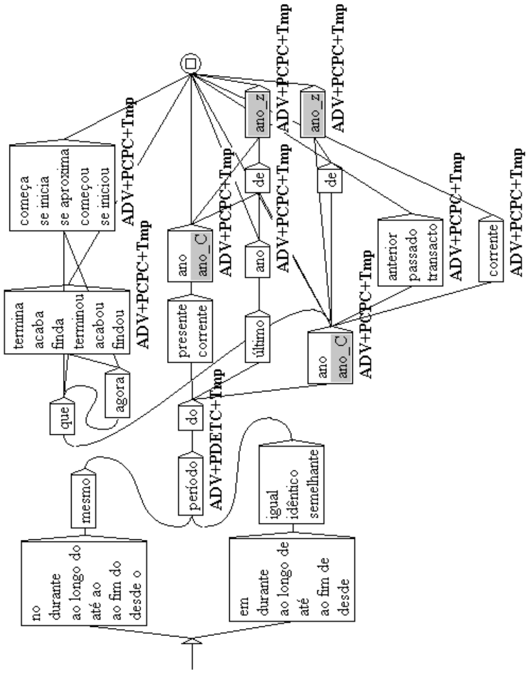

Fig. 4. *EmIgualPeríodoDoAnoModif.grf*

2 *Haver Det ano Modif*

There is a remarkably large family of adverbs that begin with the impersonal verb *haver* (there be), and express a past location in time:

(1) *<A Ana fez isso> há dois anos*
(*<Ana did that> there is two years ago*)

Like other ordinary adverbs, these expressions present some mobility in the sentence and are likely to be related with sentences like [8]:

(2) *Há dois anos que a Ana fez isso*
(*There is two years ago that Ana did that*)

It is not possible to insert adjectival modifiers. A facultative adverb, *atrás* (ago), may be added at the end of the phrase, without changing the meaning of the expression [9].
In order to describe this family of expressions, we distinguished those where the determiner is a numeral (*Num*) from those presenting other determiners.

2.1 Determinant is not a numeral

In this case, it is possible to draw up the full list of these determiners (Fig. 5):

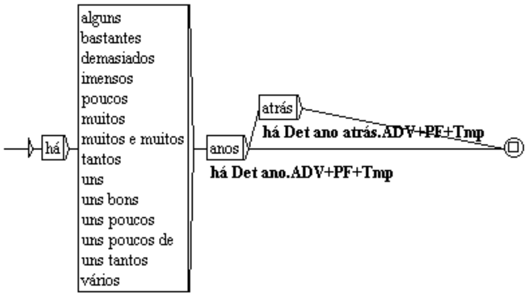

Fig. 5. *HaverDetAno_0.grf*

Prepositions *até* (until) and *desde* (since) can be inserted at the beginning of the phrase in order to express different aspectual values.

A finite set of adverbs can also be inserted in this general structure in order to express different degrees of precision (Fig. 6):

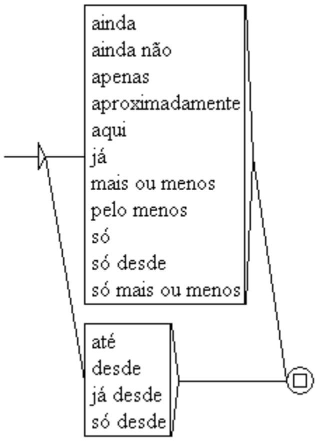

Fig. 6. *Adverbs and prepositions*

Only some of the adverbs shown at the beginning of the phrase, v. g. *apenas* (just) and *já* (already) may appear before the determiners. The choice of these adverbs depends on the determiner. There are also new determinative adverbs, *cerca de* (around), *coisa de* and *qualquer coisa como* (something like) that can now be found in this position (Fig. 7):

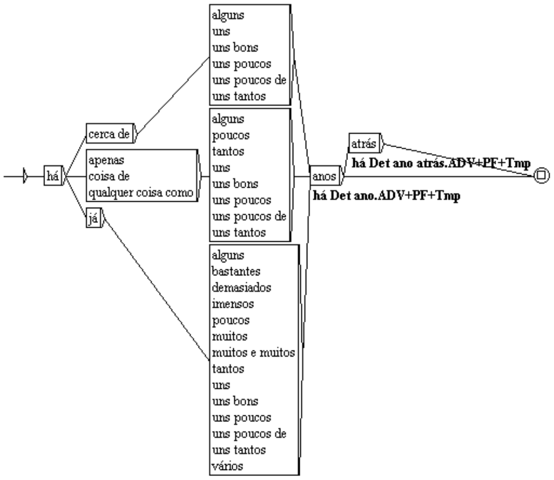

Fig. 7. *HaverDetAno_2.grf*

Finally, the set of adverbs appearing at the end of the expression is also different from the previous cases and the choice depends again on the determiners involved (Fig. 8):

[8] See M. Gross (1986: 262-265) for an extensive discussion of equivalent constructions in French. Sentences like (2) were not dealt with in this paper.

[9] With initial prepositions *de* (from) and *desde* (since) one finds other adverbial expressions such as *para cá* and *a esta parte* (up to/until now): *<A Ana tem vindo a fazer isso>* (*de* + *desde*) *há dois anos* (*para cá* + *a esta parte*) (*<Ana has been doing that> since there is two years until now*). These were represented by a separate set of graphs.

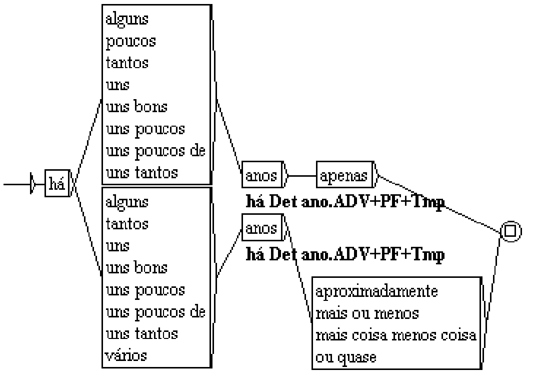

Fig. 8. *HaverDetAno_3.grf*

2.2 Determinant is a numeral

Obviously, it would not be possible to list all numerals. An available local grammar of Portuguese numerals (Ranchhod *et al.* 1999) was adapted, distinguishing those that appear just before the noun – *NumDadj*, from those that connect to the noun by the preposition *de* (of) – *NumDnom*; these grammars also recognize combinations of numbers and numerals:

<Isso aconteceu> há (*um* + *doze* + *cem* + *mil* + *30 mil*) *anos* (E + *atrás*)
(*<That happened>* (*one* + *twelve*+ *a hundred* + *a thousand* + *30 thousand*) *years ago*)

<Isso aconteceu> há (*uma dúzia* + *uma centena* + *um milhar*) *de anos* (E + *atrás*)
(*<That happened>* (*a dozen* + *a hundred* + *a thousand*) *years ago*)

The basic *há Num ano* expressions were represented by the following FST (Fig. 9):

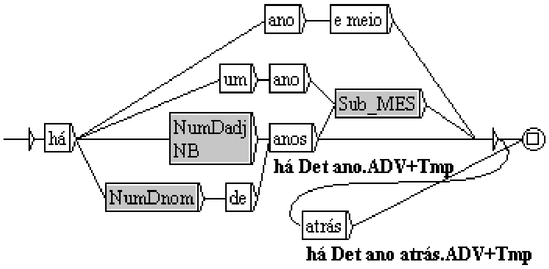

Fig. 9. *HaverNumAno_0.grf*

For numbers proper, another simple grammar was made (**NB**). Finally, the noun *ano* can be coordinated with sequences of *Num mês* (mounth) and *Num dia* (day), to express dates that are even more precise. These combinations were represented in graph **Sub_MES** (Fig. 10)[10]:

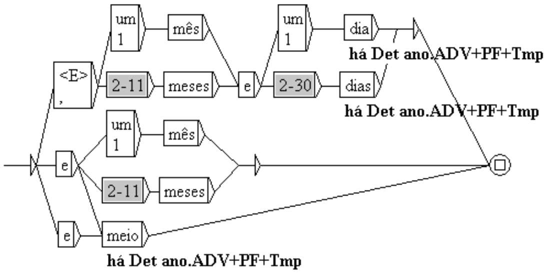

Fig. 10. *Sub_MES.grf*

As in expressions where the determiner is not a numeral, the same prepositions and adverbs can be found introducing these forms, but now the adverbs *precisamente* (precisely) and *exactamente* (exactly) can also be inserted. Similar constraints as those shown earlier between adverbs and determiners also apply in these forms.

2.3 Extension to other related families

The description of *há Det ano* adverbs can be extended to other *Ntmp*: *tempo* (time), *século* (century), *década* (decade), *mês* (mounth), *semana* (week), *dia* (day), *hora* (hour) and so on. Still, one must be careful not just to add new *Ntmp* to the existing FST. For instance, the choice of frozen modifiers depends on the Ntmp involved. Such is the case of the adjective *vindouro* (forthcoming/to come), which only combines itself with some of these *Ntmp*:

<Isso acontecerá> nos (*tempos* + *séculos* + ?*décadas* + ?**meses* + **semanas* + **dias* + **horas*) *vindouros*/*-as*)
(*<That will happen> in the times / centuries / decades / mounth / weeks / days / hours to come*)

There is also a similar family of adverbs where one finds an impersonal construction of the verb *fazer* (to do)[11] instead of *haver* (there be); these expressions may involve the presence of another adverb; this adverb can be reduced when *agora* (now) or *hoje* (today) is implicit:

<A Ana fez isso> faz (E + *hoje*) *sete dias*
(*<Ana did that> does today seven days*)

[10] Subgraphs **2-11** and **2-30** represent both the set of numerals and the corresponding numbers.

[11] M. Gross (1986: 262-265).

This family of adverbs can be represented using or adapting most of the graphs already available for the FST describing *haver Det ano*.
In fact, many of the graphs and auxiliary graphs presented in sections §1. and §2 can be reused or adapted to represent other families of expressions. Accumulative description of many more combinations can be envisaged if the methodology described here is pursued further.

3 Special combinations: *década* (decade) and *século* (century)

Several word combinations deserved special attention. This is the case of the expressions referring to *década* (decade), where a limited range of year values (from 20 to 90, but not 10) can appear:

<*Isso aconteceu*> *nos anos* (**10* + *20* + *30* + ... + *90*)
(<*That happened*> *in the years* (*ten* + *twenty* + *thirty* + … *ninety*)

The same constraints can be found with the *Ntmp decada*:

<*Isso aconteceu*> *na década de* (**10* + *20* + *30* + ... + *90*)
(<*That happened*> *in the decade of* (*ten* + *twenty* + *thirty* + … *ninety*)

With both *Ntmp,* we find similar prepositions and determiners, including partitives such as shown in Fig. 3, above. Year values can also be given in ranges:

<*Isso aconteceu*> *nos anos* (*20 –30* + *20/30*)
(<*That happened*> *in the years* (*twenty - thirty*)

Furthermore, these expressions can be expanded by indicating the century:

<*Isso aconteceu*> *nos anos 20 do século XIX*
(<*That happened*> *in the years twenty of the* XIX^{th} *century*)

Obviously, other *year – century* combinations can also be found, not referring to decades:

<*Isso aconteceu*> *nos primeiros anos do século XIX*
(<*That happened*> *in the first years of the* XIX^{th} *century*)

All these forms were represented by a family of FST, to which *Prep século* (century) combinations were added. The resulting FST present all these combinations (Fig. 11 to 13):

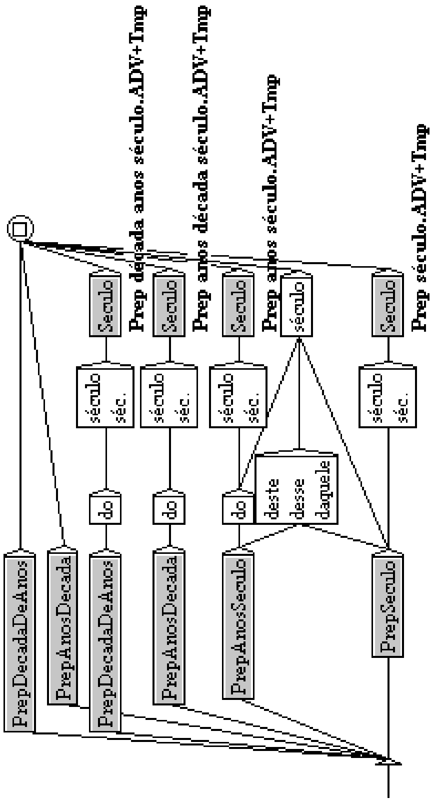

Fig. 11. *AnosDecadasSeculos.grf*

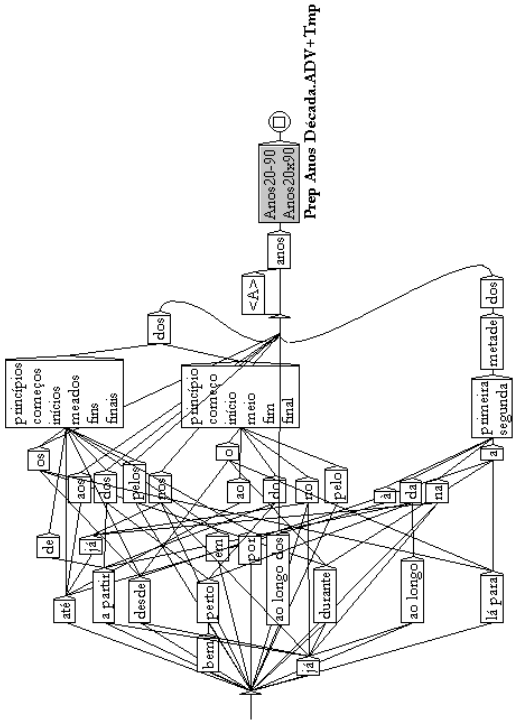

Fig. 12. *PrepAnosDecada.grf*

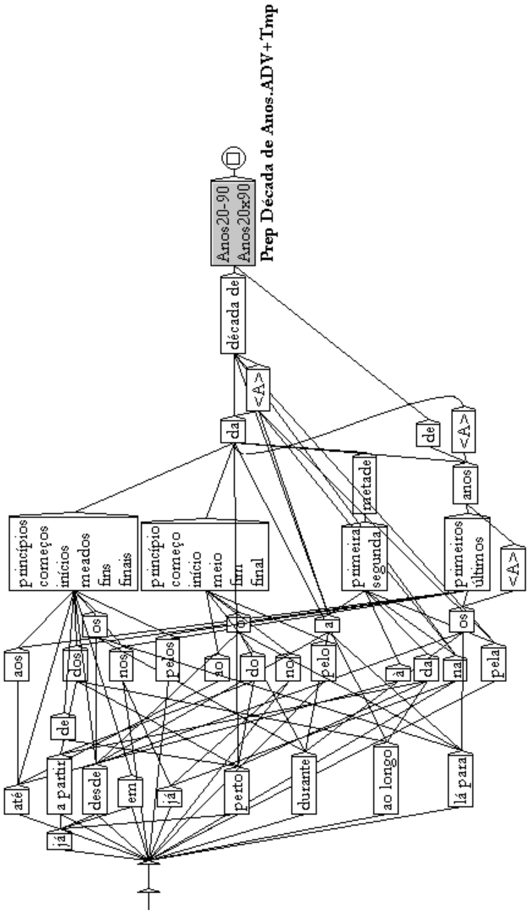

Fig. 13. *PrepDecadaDeAnos.grf*

The auxiliary graph **Seculo** includes both the roman (**NumRom**) and the ordinal numerals (**NumOrd**); the *Ntmp século* can also present several modifiers indicating the Christian or Muslim age (Fig. 14):

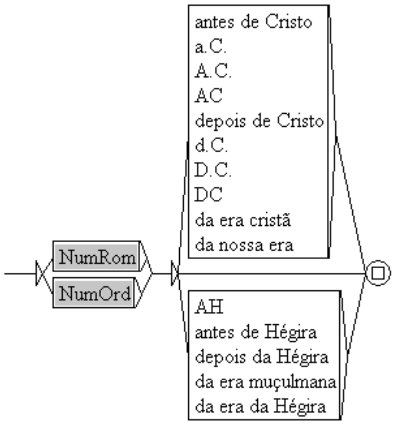

Fig. 14. *Seculo.grf*

4 *Some problems*

On one hand, the FST only partially recognize certain adverbial phrases. This occurs, for instance, when the modifier selects a prepositional phrase:

<Isso aconteceu> no ano (*anterior* + *seguinte*) *ao nascimento da Ana*
(*<That happened> in the year before/after Ana's birth*)

Certain determiners also allow such phrases:

<Isso aconteceu> no primeiro ano do reinado de Luís XIV
(*<That happened> in the first year of the kingdom of Louis XIV*)

These longer sequences represent a limitation to the simple finite-state methods used here and lower success rate considerably. On the other hand, there are certain expressions that allow the (facultative) insertion of relatively free adjectives:

<Isso aconteceu> no (*famigerado* + *glorioso* + *distante* + *remoto* + *longínquo* +) *ano de 1974*
(*<That happened> in the* (*ill-famed* + *glorious* + *distant* + *remote* + *far-away*) *year before/after Ana's birth*)

For adjectives such as *distante* (distant), *remoto* (remote) and *longínquo* (far-away), it seems possible to draw up comprehensive lists, but the same is not possible with such free modifiers as *famigerado* (ill-famed) or *glorioso* (glorious). In the FSTs built thus far, the possibility of inserting free adjectives was represented [12].

5 *Some results*

Up to now, a large variety of combinations with *Ntmp ano* have been described. It is difficult to report precise figures because local grammars involving numerical values (**NumOrd**, **NumDadj**, **NumDnom** and year values) generate an overwhelming number of combinations. If we disregard those grammars, over 40,000

[12] However, due to the caracteristics of the system, lexical FST can only tag sequences of forms (tokens), so that such expressions with free elements, represented by their categories, in this case <A> for adjectives, are not tagged. However, the FST can be used successfully to locate such expressions in texts.

different expressions with *Ntmp ano* alone have been represented by lexical FST so far and all the combinations have not been studied yet. These FST allow the identification of over 7,200 (1260 different) expressions in the working corpus – about 30 % of the occurrences of *ano*. Average recall and success rate (or precision) are relatively high (the FST retrieves about 80 % of the correct expressions found in the corpus and more than 96 % of retrieved forms are correct), but they vary depending on the family of adverbs, so that they are higher with longer sequences and lower with short expressions.

Conclusion

Complex temporal expressions play an important role in the structuring of the information in texts, especially in narrative informative discourses, where they appear with high frequency. The overwhelming number of expressions arising from their internal formal variation should not be underestimated. A detailed description of their formal variation and the corresponding combinatorial constraints, as illustrated in this paper, seems to be necessary before further processing steps can be made efficiently.

Acknowledgements

Research for this paper was partly supported by the Fundação para a Ciência e a Tecnologia (Project-grant POSI/PLP/34729/99). The Author expresses his gratitude to James Brooker and Merja de Mattos-Parreira for their help with the English version of this paper.

References

Baptista J. (1999) *Manhã, tarde, noite. Analysis of temporal adverbs using local grammars*. Seminários de Linguística 3, Universidade do Algarve, Faro, pp. 5–31.

Baptista J. and D. Català Guitart (2002), *Compound Temporal Adverbs in Portuguese and in Spanish*, In Ranchhod, E. and N. Mamede, eds., (2002) pp. 133–136.

Eleutério S., E. Ranchhod, H. Freire and Baptista, J. (1995) *A System of Electronic Dictionaries of Portuguese*. Lingvisticae Investigationes 19, pp. 157–82.

Gross M. (1982) *Une classification des phrases figées du français*, Revue québécoise de linguistiques 11-2, Presses de l'Université du Québec à Montréal, Montréal, pp. 151–185.

Gross M. (1986) *Grammaire transformationnelle du français.3 - Syntaxe de l'adverbe*, ASSTRIL, Paris, 670 pp.

Gross M. (1997) *The Construction of Local Grammars*. In "Finite State Language Processing", Y. Schabes and R. Roche, eds., MIT Press/Bradford. Cambridge/ London, pp. 329–354

Gross M. (2001) *Construção de gramáticas locais e autómatos finitos*. In E. Ranchhod, org, (2001), pp. 91–131.

Martinez-Barco P., E. Saquete and R. Muñoz (2002), *A Grammar-Bases System to Solve Temporal Expressions in Spanish*, In E. Ranchhod and N. Mamede, eds., (2002), pp. 53–62.

Maurel D. (1990), *Adverbes de date: étude préliminaire à leur traitement automatique*, Lingvisticae Investigationes 14–1, pp. 31–63.

Maurel D. (1992) *Reconnaissance automatique d'un groupe nominal prépositionnel. Exemple des adverbes de date*, Lexique 11, pp. 147–161

Molinier Ch. and F. Levrier (2000). *Grammaire des Adverbes. Description des formes en* –ment, Genève, Droz, 527 pp.

Ranchhod E., C. Mota, and J. Baptista (1999) *A Computational Lexicon of Portuguese for Automatic Text Parsing*. In "SIGLEX'99: Standardizing Lexical Ressources. 37th Annual Meeting of the ACL", College Park, Mariland, USA, pp. 74–81.

Ranchhod E. (2001) *O uso de dicionários e de autómatos finitos na representação lexical das línguas naturais*, In E. Ranchhod, org, (2001), pp. 13–48.

Ranchhod, E., org (2001) *Tratamento das Línguas por Computador. Uma Introdução à Linguística Computacional e suas Aplicações*. Caminho, Lisboa.

Ranchhod, E. and N. Mamede, eds. (2002) *Advances in Natural Language Processing*. Lecture Notes in Artificial Inteligence 2389, Berlin, Springer.

Silberztein M. (1993), *Dictionnaires électroniques et analyse automatique de texts. Le système INTEX*, Masson, Paris, 234 pp.

Silberztein M. (2000), *INTEX Manual*. ASSTRIL, Paris. http://www.bestweb.net/~intex/downloads/Manual.pdf.