Evaluation of Finite-State Lexical Transducers of Temporal Adverbs for Lexical Analysis of Portuguese Texts[*]

Jorge Baptista[1]

[1] Universidade do Algarve, Faculdade de Ciências Humanas e Sociais,
Campus de Gambelas, P-8005-139 Faro
jbaptis@ualg.pt
http://w3.ualg.pt/~jbaptis

Abstract. This paper evaluates the structure and the performance of lexical finite-state transducers (FST) used to describe and tag complex multiword temporal adverbs in texts. First, a quick overview of the formal variation allowed by expressions involving the time-related noun *ano* (year) will be presented. Results from the application to a corpus will then be presented and discussed. Finally, the paper suggests a measure for evaluating linguistic coverage and adequacy of the FST.

1 Introduction

For some time now, electronic dictionaries of both simple and compound words have been created and some are even publicly available [5,13,14,15][1]. There are, however, multiword linguistic expressions that present complex combinatorial constraints, while remaining semantically transparent. This is the case of many temporal expressions involving dates, time-related nouns and adverbs [6]. They constitute an important part of the meaning of texts, especially in informative-narrative discourses, such as newspapers. Description of these expressions is still far for complete. It is not reasonable to represent many of these complex linguistic expressions by means of electronic dictionaries because, even if they follow quite strict combinatorial rules, the sheer number of different combinations involved would make the task unfeasible. Due to their modularity, a finite-state approach seems better adequate for describing many of these expressions [7,8]. These methods have been put in place for several years and for different languages [1,2,3,4,10,11,12]. However, only limited evaluation of them or of their application to large corpora seems to have been made.

Building lexical resources for processing natural language texts is a time and effort-consuming task. Evaluation of these linguistic resources is a natural step

[*] Research for this paper was partly supported by FCT (project grant POSI/PLP/34729/99).

[1] http://www.linguateca.pt/recursos

towards public distribution, even if several methodological issues on how to evaluate them may yet remain insufficiently defined.

This paper reports on on-going research on complex temporal adverbial phrases in Portuguese. It will evaluate the performance of a small library of lexical finite-state transducers describing several families of multiword temporal adverbs built around the time-related noun (*Ntmp*) *ano* (*year*). Lexical FST are used to identify and tag these linguistic expressions in texts. Finally, the paper suggests a measure for evaluating linguistic coverage and adequacy of the FST and raises several problems concerning their accumulation, maintenance and technical limitations in their application to corpora.

2 Some Families of Multiword Temporal Adverbs in Portuguese

Many temporal adverbs are formed around time-related nouns (*Ntmp*)[2]:

segundo (second), *minuto* (minute), *hora* (hour), *dia* (day), *semana* (week), *ano* (year), *século* (century); *manhã* (morning), *tarde* (afternoon), *noite* (night/evening); *momento* (moment), *instante* (instant), *tempo* (time); an so on.

Sometimes, an entire family of adverbs can be found built around the same *Ntmp*. For example, here are some compound adverbs formed with the *Ntmp instante* (instant):

(*neste* + *nesse* + *naquele*) *instante* (at this/that instant), *de um instante para o outro* (in an instant), *a dado instante* (at a certain instant), *a todo o instante* (any instant now).

Systematic description of temporal adverbs can be done by exploring the combinations in which a given *Ntmp* enters. For each *Ntmp*, it can be structured by adopting a taxonomical approach, based on the initial preposition, the set of determinants and eventual modifiers. Following this methodology, several families of complex multiword temporal adverbs have been described so far in Portuguese [2,3,4] using *INTEX* linguistic development platform [15,16][3]. A corpus of Portuguese journalistic text – the *CETEMPúblico* corpus [4] – was used as a source for raw data. It is clear that the descriptions already available for *Ntmp ano* could be easily extended to adverbs involving other time-related nouns, probably with some minor adjustments.

So far, formal descriptions have been built of the following families of adverbs:

(1) ***daqui a Det ano***: *daqui a* (*dois* + *imensos*) *anos* (from here to two/many years, two/many years from now);

(2) ***de Det ano a Det ano***: *de ano a ano* (from year to year, each year that goes by);

[2] Notations: expressions inside brackets (…) and separated by the plus sign '+' can commutate in the given syntactic position (or not if marked with the unacceptable sign '*'). The '*E*' symbol stands for the empty string in commutation. *Adj*=adjective; *Det*=determinant; *Prep*=preposition; *Modif*=modifier. A literal translation of the examples is given to illustrate syntactic or lexical phenomena but its acceptability is irrelevant for the purposes of this paper. When necessary, an approximate translation may also be provided.

[3] http://www.bestweb.net/~intex.

[4] http://cgi.portugues.mct.pt/cetempublico. Only the first part of this corpus was used. It consists of a 58,7 Mb text and contains 9,6 million words.

de dois em dois anos (from two in two years, each two years);

(3) ***de Det ano em diante***: *deste ano em diante* (from this year onward);

(4) ***dentro de Det ano***: *dentro de alguns anos* (in some years, some years from now);

(5) ***durante Det ano***: *durante* (aquele + *o corrente* + *o presente* + *o último* + *todo o* + *próximo*) *ano* (during this/last/next year); *durante* (E + *muitos* + *demasiados* + *vários*) *anos* (during E/many/too many/several years)

(6) ***em Det ano***: *no próximo ano*, *no ano seguinte* (in the next/following year);

(7) ***Prep Det período de Det ano Modif***: *em igual período do ano passado* (in the same period of the previous year);

(8) ***num ano Modif*** : *num ano como este* (in a year such as this)

(9) (***E*** + ***Prep***) ***este ano***: *este ano* (this year); (*ao longo de* + *durante* + *até*) *este ano* (during/until this year); (*no princípio* + *em meados* + *bem perto do final* + *na primeira metade*) *deste ano* (in the begining/in the middle/very near the end/in the first half of the year).

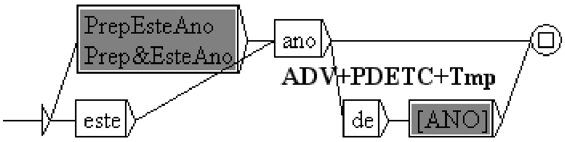

Fig. 1. ***EsteAno.fst***. In the graphic representation of FST adopted by INTEX, states are left implicit, the input symbols (words) are inside the nodes and the output symbols (linguistic information) are below the nodes: in this case, **ADV(**=adverb) is the grammatical category, **PDETC** is the formal class of compound adverbs, based on the classification given in [6], and **Tmp** (=time) is a semantic feature. Auxiliary graphs appear in gray boxes: **PrepEsteAno** and **Prep&EsteAno** represent different combinations of prepositions and adverbs with determinant *este*; **[ANO]** describes several formats for year numbers.

(10) ***Det ano*** (***atrás*** + ***antes*** + ***depois***) [5]: (*dois* + *vários* +*muitos*) *anos* (*atrás* + *antes* + *depois*) (two/several/many year ago/before/after)

(11) ***por Det ano***: *por* (*dois* + *muitos e longos*) *anos* (for two/many long years);

(12) (***a*** + ***em*** + ***por***) ***cada ano*** (E + ***que passa***) : *em cada ano* (each year);

(13) (***E*** + ***Prep*** + ***Adj***) ***todos Det anos*** : *todos os anos* (every year); *ao longo de/após todos esses anos* (during/after all those years); (*decorridos* + *passados* + *volvidos*) *todos esses anos* (after all those years);

(14) ***em Det anos*** [20-90]: *nos anos* (*90* + *noventa*) (*in the 90's* + *nineties*) [6]

(15) ***fazer Det ano***: *faz* (*dois* + *muitos*) *anos* (two/many years ago)

[5] In some cases, expressions of one family overlap those of another. For instance, most adverbs from this family are included in *haver Det ano* expressions (16).

[6] This family makes part of a larger set, which also involves *Ntmp década* (decade) and *século* (century). These expressions were described together for having in common the finite set *20, 30... 90* (and the corresponding numerals). Combinations like **nos anos* (*1900* + *1910*) are unacceptable in Portuguese.

(16) (**E** + ***Prep***) ***haver Det ano***: *há* (*dois* + *muitos*) *anos* (two/many years ago)

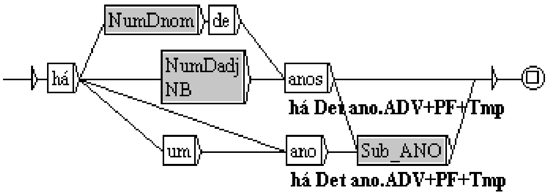

Fig. 2. ***HaverNumAno00.fst***. This is one of the FST of the *haver Det ano* family. In this FST, ***há Det ano*** functions as a conventional lemma, **PF** is the formal class of compound adverbs. Auxiliary graphs: **NB** represents numbers while **NumDnom** and **NumDadj** represent numerals; **Sub_ANO** describes complements with *Ntmp mês* (months) and *dia* (day).

3 Evaluating Lexical Finite-State Transducers

It is possible to conceive two basic measures for evaluating lexical finite-state transducers as those illustrated above. The first and currently most used measures have to do with the **performance** of the FST, i.e., with results obtained from applying FST to corpora. The second has to do with the **lexical coverage** and **linguistic adequacy** of the FST, i.e., how many correct, different expressions can it potentially recognize.

3.1 Methods

Evaluating the performance of FST when they are applied to a corpus involves two basic measures: **recall** and **success** rate.[7] In order to calculate recall, it is necessary to determine the number of correct expressions in the corpus (this could be called the **optimal value**). If the corpus were relatively small, it would then become possible to determine this value. Obviously, this task has to be done manually so optimal value is not always available. In some cases, it may be possible to devise strategies to obtain the optimal value.[8] From these basic measures it is possible to obtain **silence** (the difference between optimal value and success rate), which corresponds to the number

[7] **Recall** can be defined [9] as the ratio of correct expressions retrieved by the FST over the total correct expressions in the corpus. **Success rate** (also called **precision**) is the percentage of correct expressions of all expressions retrieved by the FST.

[8] Sometimes, it is possible to build a finite-state automaton (FSA) just in order to extract a subset of the corpus where all expressions being described *must* be present. The total number of occurrences of the pattern described by this FSA would be the **universe** of potential application of the FST. If the expressions found by the FST were removed from this set, the remaining matches may be sufficiently small for manual scrutiny, thus yielding **silence**. This was the methodology used in this paper.

of linguistic expressions that the FST fails to recognize, and **noise** (or **failure rate**), i.e., the number of sequences found by the FST that do not correspond to the linguistic expressions one wanted to describe. Ideally, silence should be null and noise as limited as possible.

Finally, because the FST were built using raw data from an initial corpus, results should be compared with those obtained from other (similar and/or different) corpora. Due to space limitations, this paper will only present results of some families of adverbs from a single corpus.

3.2 Results

Results vary considerably depending on the family of adverbs. Some FST are still under construction, hence their results are not satisfactory yet.

Table 1. Performance of FST on corpus *CETEMPúblico Part 01*. The first five columns indicate brut results: D=dictionary is the number of different compound lexical entries found in the corpus; U=universe; OV=optimal value; L=locate is the number of expressions retrieved by the FST; CL=correct locate is the correct number of expressions retrieved by the FST; R=recall (CL/OV); SR=success rate (CL/L); N=noise (L-CL/L); S=silence (OV-CL)/OV).

FST	**D**	**U**	**OV**	**L**	**CL**	**R**	**SR**	**N**	**S**
(1) *daqui a ano*	66	144	143	143	137	95.80	95.80	4.20	4.20
(2) *de ano a ano*	9	39	35	35	35	100,00	100,00	0,00	0,00
(3) *de ano em diante*	2	2	2	2	2	100,00	100,00	0,00	0,00
(4) *dentro de Det ano*	33	154	151	123	120	79.47	97.56	2.44	20.53
(5) *durante Det ano*	253	1036	943	699	654	69.35	93.56	6.44	30.65
(9) *Prep este ano*	51	2526	1978	1981	1806	91.30	91.17	8.83	8.70
(11) *por Det ano*	33	1189	235	162	159	67.66	98.15	1.85	32.34
(13) *todos anos*	26	260	257	263	243	94.55	92.40	7.60	5.45
(14) *Prep ano* [20-90]	219	-	-	832	813	-	97.72	2.28	-
(16) *haver ano*	567	3823	2980	2964	2964	99.46	99.70	0.30	0.54
Total/Average	1259	9173	6724	7213	6933	78.80	96.61	3.39	11.38

3.3 Discussion

Results of some FST (1-3, 9, 13 and 16) show very high recall and success rate. This has to do with the fact that some multiword expressions are relatively long and complex so that they only rarely give rise to formal ambiguity with other word combinations. For instance, the low noise value (0.30 %) observed with *haver ano* results only from expressions with adverb *aqui* (here), such as (all examples were taken from the corpus):

As suas tropas estão ***aqui há anos***
(His troups are here for years now)

Here, adverb *aqui* is a spatial locative complement of the verb, followed by the adverb with *haver-ano*, which is a time locative on the entire sentence. However, this adverb actually exists in the corpus:

Aqui há anos *a filosofia era outra*

(Years ago the philosophy was different)

Considering this strictly local ambiguity, the FST should signal these expressions as potentially faulty. On the other hand, low silence was due to expressions where *haver* is at the indicative imperfective past tense (*havia*).

[...] *deixado de ser capital* ***havia apenas dois anos***

(was no longer the capital for two years only)

These expressions are relatively rare in the corpus. They had not been considered during the construction of the FST. It is relatively easy to add them to the FST. In other cases (5 and 11), low recall happens when some preposition, adverb or modifier were not included in the FST yet. This is a matter of completion of the FST. It should be noted, however, that success rate is high.

Another problem arises from the overlap of two formally similar adverbs. As noted before, most expressions of (10) are included in the description of (16), thus they were not included in Table 1. The formal definition of year values (sequences of four or two digits: *1993* or *93*) give rise to significant overlap because the FST retrieves *1993* twice: *1993* and *19*, the second being obviously incorrect. This situation reduces recall and precision in a significant way but could be solved by imposing to the system to retrieve only the longer sequence.

Finally, it is difficult to calculate recall for some FST. In order to retrieve the universe of adverbs such as (14) *em Det ano*, the regular expression[9]: `em (<MOT> + <NB>)* (ano+anos)` produces over 200,000 expressions from the corpus, too many to be manually verified[10]. Therefore, only success rate was presented.

3.4 Lexical Coverage and Linguistic Adequacy

When evaluating lexical FST, it is also possible to consider another measure: how many different expressions the FST represents? Are they all correct? This requires generating all the linguistic expressions described by the FST and verifying if they are correct, so that ideally the FST would only represent combinations authorized by the language.

It also involves deciding what it is meant by “different” expressions. For instance, what should be the status of productive subsets of linguistic expressions (e.g. numbers, numerals) or otherwise recurrent close sets (days of the week, months) in this counting? To illustrate the issue, consider the auxiliary graph **`Sub_ANO`** appearing in Fig. 2. This automaton describes eventual complements of the *Ntmp*, expressing the exact number of months and days. If all variation regarding numeral and number

[9] The regular expression is given in the Intex format: <MOT> and <NB> are in-built symbols that stand for any word or any number, respectively; ‘*’ is the Kleene operator.

[10] A sampling procedure could be used instead.

determinants were generated, it would produce 2,663 different combinations. If not, only 7 different strings should be counted [11]:

<há um ano> (*e meio* + *e* [1] *mês* + *e* [2-11] *meses* + , [1] *mês e* [1] *dia* + , [1] *mês e* [2-30] *dias* + , [2-11] *meses e* [1] *dia* +, [2-11] *meses e* [2-30] *dias*)

Probably, these finite sets should not count. Nevertheless, how would one decide where should these limits be imposed? Should even complements such as those above be counted or discarded? Should the same be done to recurrent sets of prepositions, determiners, adverbs and modifiers? It is difficult to have clear answers. Provisory, only numerals and some limited sets have not been counted. Even so, the (16) *haverDet ano* family alone generates more than 9,000 different expressions; the (5) *durante Det ano* family is even more impressing: over 31,000 different expressions.

Consider now the issue of linguistic adequacy. A simple measure could be defined as the percentage of correct expressions generated by the FST. For example, the FST for the (9) *Prep este ano* family generates 424 different expressions. It is necessary to verify each string manually in order to determine if there is over-generation of incorrect forms (fortunately, this was not the case here). However, verifying the results of the language generated by some FST may not be so simple, as in the case of families (5) and (16) above. The number of (correct) expressions generated by a FST might be compared with silence, thus giving an approximation to the notion of linguistic coverage, if the corpus is considered to be representative for the phenomena under study.

In spite of this, for some FST it may be convenient, for simplicity purposes, to loose the formal constraints on word combinations, otherwise the set of FST would be too complex to manage and maintain. Over-generation may not necessarily affect success rate, since clearly incorrect strings will probably not occur in texts. However, linguistic adequacy may be reduced.

4 Final Remarks

The size and complexity of complex multiword temporal adverbs should not be underestimated. They constitute a non-trivial challenge to computational processing (consider, for instance, the resolution of temporal references [10]). Such large FST may also render significantly slow the lexical processing of texts.

In order to ensure efficient maintenance of cumulative data, the description needs to be structured with clear taxonomical principles, considering the overwhelming size and significant overlap of some families of expressions. The compromise between linguistic adequacy and efficiency in results should be taken into consideration when evaluating linguistic resources such as the lexical transducers described here.

[11] Numbers inside square brackets stand for ranges of both numerals and numbers.

References

1. ACL Workshop on Temporal and Spatial Information Processing. Toulouse, France (2001)
2. Baptista, J.: Manhã, tarde, noite. Analysis of temporal adverbs using local grammars. Seminários de Linguística 3 (1999) 5–31
3. Baptista, J., Català-Guitart, D.: Compound Temporal Adverbs in Portuguese and in Spanish. In: Ranchhod, E., Mamede, N. (eds.): Advances in Natural Language Processing. Lecture Notes in Computer Science, Vol. 2389. Springer-Verlag, Berlin Heidelberg New York (2002) 133-136
4. Baptista, J.: Some Families of Compound Temporal Adverbs in Portuguese. In: Laporte, E. (org.): Workshop on Finite-State Methods in Natural language Processing at EACL'2003 (to appear).
5. Eleutério, S., Ranchhod, E., Freire, H., Baptista, J.: A System of Electronic Dictionaries of Portuguese. Lingvisticae Investigationes 19 (1995) 157–182.
6. Gross, M.: Grammaire transformationnelle du français.3 - Syntaxe de l'adverbe, ASSTRIL, Paris (1986) 670 pp.
7. Gross, M.: The Construction of Local Grammars. In Schabes, Y., Roche, E. (eds.): Finite State Language Processing. MIT Press/Bradford. Cambridge/ London (1997) 329–354
8. Gross, M.: Construção de gramáticas locais e autómatos finitos. In: Ranchhod, E. (org.): Tratamento das Línguas por Computador. Uma Introdução à Linguística Computacional e suas Aplicações. Caminho, Lisboa (2001) 91–131
9. Hirschman, L., Mani, I: Evaluation. In: Mitkov, R. (ed.): The Oxford Handbook of Computational Linguistics. Oxford University Press, Oxford (2003) 414-429
10. Martinez-Barco, P., Saquete, E., Muñoz, R.: A Grammar-Bases System to Solve Temporal Expressions in Spanish. In: Ranchhod, E., Mamede, N. (eds.): Advances in Natural Language Processing. Lecture Notes in Computer Science, Vol. 2389. Springer-Verlag, Berlin Heidelberg New York (2002) 53-62.
11. Maurel, D.: Adverbes de date: étude préliminaire à leur traitement automatique. Lingvisticae Investigationes 14–1 (1990) 31–63.
12. Maurel D.: Reconnaissance automatique d'un groupe nominal prépositionnel. Exemple des adverbes de date. Lexique 11 (1992) 147–161
13. Ranchhod, E.: O uso de dicionários e de autómatos finitos na representação lexical das línguas naturais. In: Ranchhod, E. (org.): Tratamento das Línguas por Computador. Uma Introdução à Linguística Computacional e suas Aplicações. Caminho, Lisboa (2001) 13-48.
14. Ranchhod, E., Mota, C., Baptista, J.: A Computational Lexicon of Portuguese for Automatic Text Parsing. In: SIGLEX'99: Standardizing Lexical Ressources. 37th Annual Meeting of the ACL, College Park, Maryland, USA (2001) 74–81.
15. Silberztein M. Dictionnaires électroniques et analyse automatique de texts. Le système INTEX, Masson, Paris (1993) 234 pp.
16. Silberztein, M.: INTEX Manual. ASSTRIL, Paris (2000). http://www.bestweb.net/~intex/ downloads/ Manual.pdf.