Developing a UNL decodifier for Brazilian Portuguese

Núcleo Interinstitucional de Lingüística Computacional (NILC)[*]

USP - São Carlos, UFSCar, Unesp - Araraquara

CP 668 - ICMSC-USP, 13560-970 São Carlos, SP, Brazil

Abstract

This paper describes results from the development of a decodifier for decoding UNL into Brazilian Portuguese, which is being carried out by NILC. The two main tasks are the creation of generation rules for feeding the DeCo system provided by the UNL Center and the development of a lexicon of Portuguese headwords that are associated with the universal words (UWs).

1. Introduction

The *Núcleo Interinstitucional de Lingüística Computacional* (NILC) is currently responsible for developing the natural language processing activities within the Universal Networking Language project (UNL-BR). The first stage in the UNL project is devoted to two tasks: i) compilation of a lexicon of headwords for each given language where each headword is associated with a specific Universal Word contained in the UW dictionary provided by the UNL Center; ii) creation of a set of generation rules that allow sentences represented in the UNL formalism to be decodified into the particular language under study.

This progress report refers to work carried out after April, 1997, when the UNL Center made the dictionary of Universal Words (UWs) and the DeCo system for decodification of UNL sentences available for European languages employing the Latin alphabet. Also provided was a corpus of 20 sentences with UNL representation which are used by all development teams in the UNL project. Before proceeding to the details on the methodologies adopted and on the results obtained so far, we provide some background information on the DeCo system and on the UW dictionary in Section 2. For the DeCo, in

[*] M.G.V. Nunes; S.M. Aluísio; A.G. Bonfante; B.C. Dias-da-Silva; R. Hasegawa; M.A.C. Jesus; R.T Martins; G. Montilha; O.N. Oliveira Jr.; L.H.M. Rino; C.R. Sossolote; C. Zavaglia.
http://www.nilc.icmsc.sc.usp.br.

particular, the description is relatively longer since this tool is the core of the decodification system. The mapping of the UWs onto Portuguese headwords is described in Section 3, while Section 4 discusses the methodology for mapping Relation Labels (RLs) and Attribute Labels (ALs) onto surface structures for Brazilian Portuguese, also including a description of the generation rules to be fed into the DeCo system. Conclusions and indication of further work are given in Section 5.

2. Background on the material provided by the UNL Center

2.1. The DeConverter Tool - DeCo

The DeCo system is aimed at decodifying any given UNL sentence into a sentence in a target language (Figure 1). It comprises basically two main processes: the resolution of semantic relationships between UWs that compose the UNL sentence, seen as a nodenet, along with the resolution of their grammatical attributes, and the control over the windows that allow for the modification of the nodelist. The nodelist holds information that is gradually modified to generate the final sentence in the target language. As such, it actually holds the result of the computation. During intermediate stages of the decodification, the nodelist expresses, e.g., headwords and affixes (to be attached to headwords in order to produce a full word).

DeCo combines a word dictionary of a language employing the Latin alphabet with a set of its corresponding generation rules, in order to provide the adequate full system to decodify any given UNL sentence into a sentence written in that particular language. In short, once a target language is given, its corresponding word dictionary and set of generation rules are the inputs to DeCo and they both complement a full decodifier for that language. The decodifier is thus a language generation system, which is used to produce the target sentence.

Dictionary entries and generation rules are provided by every client (i.e., a UNL team) working on a special target language. Once a UW is chosen to be processed, the dictionary is looked up before the application of any generation rule, once a UW is chosen to be processed. The goal of looking up an entry is to associate the node information to a UW. For example, its headword, the corresponding UW, and its grammatical attributes. Once information corresponding to the UW at hand is fully retrieved, DeCo begins searching for applicable generation rules, in order to proceed to the construction of the target sentence. In this way, chosen generation rules are those more likely to allow for the derivation of a grammatical sentence in the target language.

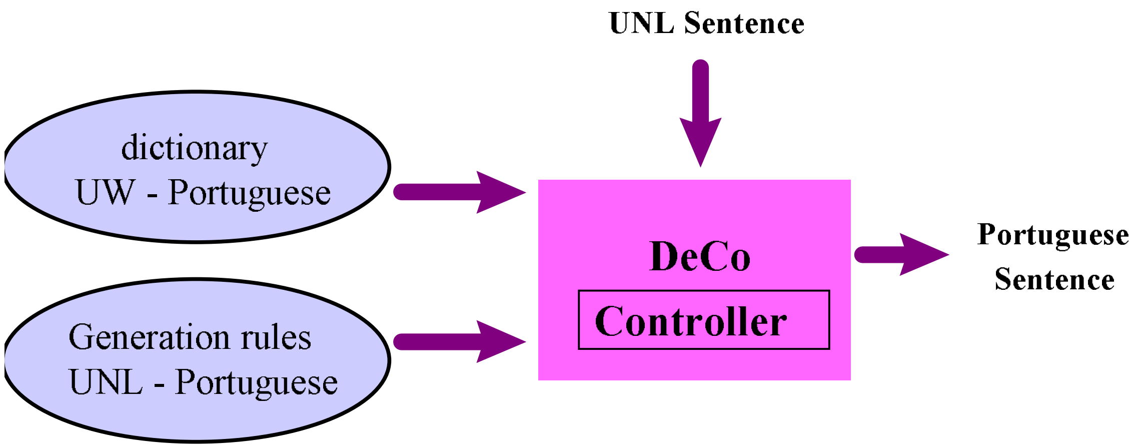

Figure 1. DeConverter Process using DeCo

Generation rules are applied only to the nodes in the nodelist. However, the status of both, the nodelist and the nodenet (for further details on these, please refer to the deconverter for Latin Languages Manual, Version 1.0) are taken into account during decodification. The generation process is carried out under the control of two types of windows: generation windows and condition windows. The left (LW) and right (RW) generation windows look into the contents of the nodelist, specially into the grammatical features of each node at the windows; the condition windows look into the neighbors of the generation windows, also in respect to the grammatical features of the nodes in question.

In checking the neighboring nodes on both sides of the generation windows, condition windows support further modifications of the nodelist, by bringing about information on the feasibility of application of a specific generation rule. Together, both types of windows provide information to look for generation rules that match the current status of the nodelist. When applied, those rules produce the following alterations of the nodelist (not necessarily exclusive): a) modification of the nodelist by updating the contents of the involved nodes (adding, or extracting, some of their grammatical attributes); b) insertion of a new node (either from the nodenet or from elsewhere) in a position that is relative to the generation windows. After an insertion or a modification, a movement of the generation windows can also happen.

Figure 2 shows the status of the nodelist and nodenet in the beginning of the decodification process of a UNL sentence containing three relation labels (agt., obj., aoj) among three UWs (people, tower, huge). When decodified into Portuguese, the UNL sentence gives rise to *"O povo construiu uma torre imensa"*.

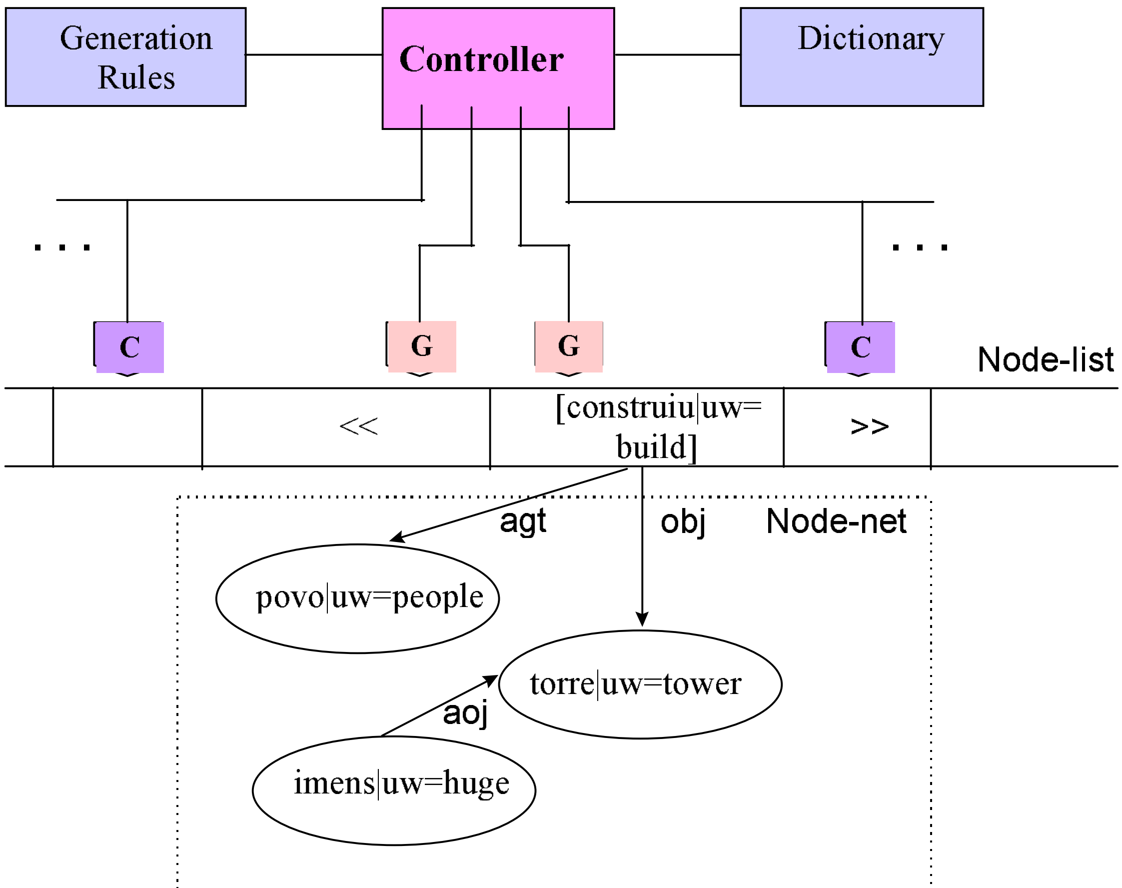

Figure 2. DeCo configuration for generating "O povo construiu uma torre imensa"

Insertion of nodes in the nodelist which are not apparent in the nodenet aims at describing special conditions of a UW, e.g., suffixes signalizing verb tense, such as "ed", or blanks (between words). This type of insertion concerns mostly the occurrence of modifiers of the nodes already occurring in the nodelist and it is signaled by means of a grammatical feature initiated by "!". Features like this trigger a special process of resolving grammatical features. For example, the token *!present* triggers the generation of the lexical item derived from a verb in the infinitive form by adding present tense features.

In general, we can summarize the application of generation rules in DeCo in the following way: the inclusion of nodes in the nodelist is chosen on the basis of a) the content of the LW and the RW; b) the neighbors of such windows; c) the node of the nodenet which is candidate to go into the nodelist. Each node of the nodenet is inserted in the nodelist and later replaced by the corresponding word to get the final lexicalized form.

According to the above, the syntax of DeCo's generation rules comprises three different patterns, as follows:

1) attribute modification:
{condition:action:relation:role}M{condition:action:relation:role}

2) insertion (of n_j) on the left of any generation window:
"condition:action:relation:role"M{condition:action:relation:role}

where *relation* links the node pinpointed by one of the generation windows, say n_i, to a node in the nodenet, say n_j, and n_i and n_j satisfy the conditions to apply the generation rule. When insertion of a node in the nodelist is carried out, the indicated actions take place.

3) insertion on the right of any generation window:
{condition:action:relation:role}M"condition:action:relation:role"

This insertion is similar to the previous, but the node from the nodenet is now inserted on the right.

Above, M can stand for L, R, : or ?, respectively representing *left* and *right* movement of both generation windows, alteration of grammatical attributes of the nodes pinpointed by the generation windows, and backtracking. A backtracking process can happen over the choice of a headword or over the generation rules (returning back to a previous state to match a different rule), but never on the insertion of a node in the nodelist. Any movement of the generation windows always happens after the application of the rule. Both windows move together, i.e., if one goes to the right, so does the other and vice-versa. Moreover, if no rewriting action occurs in the node of the LW, the position of the generation windows remains unchanged. In general, the following are the possible movements of the generation windows:

1) When M = ":" (attribute change)
• If no rewriting action occurred in the node at the LW, there is no movement.
• If a rewriting action occurred in the node at the LW, there is a movement to the left.

2) When M = "R" (right shift)
• The generation windows move to the right.

3) When M = "L" (left shift)

• The generation windows move to the left.

The association of information provided by the condition and the generation windows allows for addressing context dependency.

2.2. The Dictionary of Universal Words

The UNL Center made available a dictionary containing approximately 220.000 Universal Words that were generated from a set of 95.000 distinct English words. The syntax of the UW dictionary is as follows:

[] "UW" ( ) < >;

where: [] contains the headword in the target language (Portuguese in our case)
"UW" defines a meaning
( ) contains grammatical features of the headword
<> brings restriction information for the headword, including priority and frequency of the headword, which are used in the decodifying and codifying processes, respectively

The entries in Figure 3 illustrate the contents of the UW dictionary. The entries for a given word in English correspond to its different meanings. It should be noted that semantic restrictions are imposed to the most generic UW (threaten, in this example), which delimits its meaning giving rise to distinct headwords in the target language. Such restrictions usually involve relation labels (e.g. agt, obj, and inclusion and synonymy as icl, equ) and other UWs (e.g. danger, human, trouble).

```
[]{} "threaten" ();
[]{} "threaten(agt>human,obj>danger)" ();
[]{} "threaten(agt>human,obj>entity)" ();
[]{} "threaten(agt>human,obj>human)" ();
[]{} "threaten(agt>human,obj>trouble)" ();
[]{} "threaten(icl>event)" ();
[]{} "threaten(icl>event,obj>human)" ();
```

Figure 3. Entries in the UW dictionary for the English word "threaten"

In mapping the UWs onto Portuguese headwords the grammatical as well as the semantic attributes must be included in a list within the parentheses on the right hand side of the UW (see Figure 3). Furthermore, information on frequency and priority are necessary for decodification and codification purposes. The final format for the UW dictionary is then:

[headword] canonical-word "UW" (grammatical features) <P,f,p>;

where: headword is the Portuguese word corresponding to the UW;
canonical-word is the non-analyzed form of the Portuguese headword;

(grammatical features) is the set of grammatical and semantic attributes for the headword;
P denotes the Portuguese language;
f and p are numbers that express the frequency and priority of the headword.

The methodology for establishing headwords for the UWs is discussed in Section 3.

2.3. The Corpus of UNL sentences

The corpus of UNL sentences has been used as a common ground to all development teams involved in the UNL project. It consists of 20 English sentences obtained from the UNL booklet [1], for which the UNL representation has been provided. Figures 4 and 5 show respectively the UNL codification for the following Portuguese sentences: (a) "*Há muito tempo, na cidade de Babilonia, o povo começou a construir uma torre imensa, que parecia alcançar os céus*"; (b) "*Ciente das intenções dos homens, Deus ficou irado e criou então uma multiplicidade de línguas, de modo que eles não pudessem mais se comunicar uns com os outros*".

3. Mapping Portuguese headwords to Universal Words

The process of ascribing Portuguese headwords to UWs from the UW dictionary started by addressing the most frequent words, in both languages, English - from which the lexical representation was retrieved for the meaning of the UWs - and Portuguese, the target language. This methodology is aimed at obtaining a set of Portuguese headwords which cover the ordinary uses of Portuguese. As for the most frequent English words, we selected 2000 words that have been quoted as the most frequent for that language [2]. Using these 2000 English words, a set of 26.000 UWs were selected for which the corresponding Portuguese headwords are being established.

Sentence (a)

```
[S]
;<1-002>Long ago, in the city of Babylon, the people began to build a huge tower, which seemed about to reach the heavens.
tim(begin(icl>event).@entry.@pred.@past, long ago)
nam(city(icl>place).@def, Babylon(icl>city))
ppl(begin(icl>event).@entry.@pred.@past, city(icl>place).@def)
agt(begin(icl>event).@entry.@pred.@past, people(icl>human).@def)
obj(begin(icl>event).@entry.@pred.@past, build(equ>construct,agt>human,obj>structure).@pred)
agt(build(equ>construct,agt>human,obj>structure).@pred, people(icl>human).@def)
obj(build(equ>construct,agt>human,obj>structure).@pred, tower(icl>building).@indef)
aoj(huge(aoj>entity), tower(icl>building).@indef)
aoj(seem(icl>event).@pred.@past, tower(icl>building).@indef)
obj(seem(icl>event).@pred.@past, reach(gol>entity,obj>entity).@pred.@begin-soon)
obj(reach(gol>entity,obj>entity).@pred.@begin-soon, tower(icl>building).@indef)
gol(reach(gol>entity,obj>entity).@pred.@begin-soon, heaven(ant>hell).@def.@pl)
[/S]
```

Figure 4.UNL sentence for "*Há muito tempo, na cidade de Babilonia, o povo começou a construir uma torre imensa, que parecia alcançar os céus*"

This work has been carried out in two ways: i) by treating the so-called closed classes of words, such as articles, prepositions, conjunctions, etc., and ii) by considering nouns and verbs. In the latter case, after ascribing Portuguese headwords to the UWs, the grammatical and semantic features are included in a later stage. It must be stressed that ascribing headwords is not simply a translation process, since the restrictions accompanying the headwords must be interpreted and taken into account accordingly. In fact, this makes it difficult to automate the process of finding Portuguese headwords for the UWs from an electronic dictionary, for instance. Nevertheless, we have been investigating ways of automatically interpreting the restrictions, so that the process can be at least semi-automated.

Sentence (b)

```
[S]
;<1-003>Aware of the people's intentions, God became angry and so He created a polyglot of languages so that they would no longer be able to communicate with each other.
pos(intention.@def.@pl, people(icl>human))
agt(aware(icl>state).@pred, god)
obj(aware(icl>state).@pred, intention.@def.@pl)
man(become(icl>event).@pred.@past, aware(icl>state).@pred)
obj(become(icl>event).@pred.@past, god)
ol(become(icl>event).@pred.@past, angry(icl>state).@pred)
and(create(icl>event).@pred.@past.@entry, become(icl>event).@pred.@past)
man(create(icl>event).@pred.@past.@entry, so(equ>therefore))
agt(create(icl>event).@pred.@past.@entry, He)
obj(create(icl>event).@pred.@past.@entry, polyglot.@indef)
mod(polyglot.@indef, language(icl>language).@pl)
pur(create(icl>event).@pred.@past.@entry, communicate(agt>human,ptn>human).@pred.@apodosis.@ability)
man(communicate(agt>human,ptn>human).@pred.@apodosis.@ability, no longer)
agt(communicate(agt>human,ptn>human).@pred.@apodosis.@ability, they(icl>human))
man(communicate(agt>human,ptn>human).@pred.@apodosis.@ability, each other)
[/S]
```

Figure 5.UNL sentence for "*Ciente das intenções dos homens, Deus ficou irado e criou então uma multiplicidade de línguas, de modo que eles não pudessem mais se comunicar uns com os outros*"

The inclusion of grammatical and semantic features may also benefit from an automatic pre-processing, especially in the case of the grammatical features that are available for more than 1.5 million words in a lexicon developed at NILC for Brazilian Portuguese. The work for the lexicon has been based on a number of dictionaries and also on the software *Wordnet* for *PC, version 1.5*.

Tests made with DeCo have shown that it is more convenient to adopt a dictionary of analyzed forms for the headwords. That is, the headwords in the dictionary will also be represented according to their roots, as shown in Figure 6. However, for nouns, adjectives and verbs, both analyzed and non-analyzed forms will be available so that words can be generated with the correct gender and number.

In the following, a simplified version of the grammatical attributes for each category of headword is presented.

```
smooth(aoj>movement)
[perfeit] {} perfeito "smooth(aoj>movement)"(stem,plural,1arg,rege(de)(em)) <P,0,0>;
communication(icl>connection)
[transmissão] {} transmissão communication(icl>connection"
(stem,^alomorfe,fem,2arg,rege(a)(para)(por)(de),deverbais,comum) <P,0,0>;
[transmissõ] {} transmissão "communication(icl>connection)"
(stem,alomorfe,plural(es),2arg,rege(a)(para)(por)(de),deverbais) <P,0,0>;
```

Figure 6. Examples of analyzed and non-analyzed lexical entries

(a) noun
structure: root or the non-analyzed form (including gender and number)
form: main root or a variant if there exists one
number: only for the non-analyzed words: sing, pl, 2N
plural formation: only for the roots: with **-s** or **-es**;
gender: masc, fem, 2G;
transitivity: zero, one or two arguments;
government agreement: list of prepositions
reference: places, persons, objects, gender;
type: common or proper.

(b)Adjective
structure: root or the non-analyzed form
form: main root or a variant if there is one
plural formation: with **-s** or **-es**;
gender: masc, fem, 2G;
transitivity: zero, one or two arguments;
government agreement: list of prepositions

(c) Personal and Possessive Pronouns
case (only for the personal pronouns);
number: sing, pl, 2N
gender: masc, fem, 2G
person: 1st., 2nd., 3rd.

(d) Undefined, Demonstrative, Interrogative Pronouns
number (for the undefinite and demonstrative):sing, pl, 2N
gender (for the undefinite and demonstrative): masc, fem, 2G

(e) Articles
number: sing, pl;
gender: masc, fem;
reference: definite, undefinite

(f) Conjunctions and Conjunctional Phrases
type: sets of the types of conjunctions

(g) Verbs
conjugation: 1st, 2nd, 3rd
regularity: regular, irregular;
aspect (for the irregular): simple past, imperfect past;
paradigms: one of nineteen paradigms;
transitivity: auxiliar, linking verb, intransitive; direct transitive; indirect transitive; bitransitive;
government agreement: list of prepositions;
reference: state, action, movement, psychological, dicendi.

(h) Numerals
gender (only for the cardinal and multiplicative): masc, fem, invariable.

(i) Prepositions and Prepositional Phrases
type: set of types of prepositions and prepositional phrases.

Figure 7 shows some UWs contained in the sentences of Figure 4 and 5.

```
[homens] {} homens "people(icl>human)" (^stem,plural,fem,0arg,pessoas,comum) <P,0,0>
[torre] {} torre "tower(icl>building)" (stem,plural(s),feminino,0arg,lugares,comum) <P,0,0>
[imens] {} imenso "huge(aoj>entity)" (stem,^alomorfe,plural(s),masc,1arg) <P,0,0>
[que] {} que "which" () <P,0,0>
[quase] {} quase "about" () <P,0,0>
[céu] {} céu "heaven(ant>hell)" (stem,plural(s),masc,0arg,lugares,comum) <P,0,0>
[ciente] {} ciente "aware(icl>state)" (stem,plural(s),2G,1arg,rege(de)) <P,0,0>
[Deus] {} Deus "God" (stem,plural(es),masc,0arg,deverbal,próprio) <P,0,0>
```

Figure 7. Examples of dictionary entries

4. Decodification Rules for Portuguese

Generation rules are being developed for decodifying the 20 UNL sentences from the corpus into Brazilian Portuguese. For that, the sentences were manually translated into Portuguese which allowed us to map the Relation and Attribute Labels onto syntactic structures for Brazilian Portuguese. For instance, the RL *agt* is usually realized as the subject of a sentence [3]. This mapping process yielded a set of generic generation rules, even though the focus was on the sentences of the corpus. Up to now, 22 RLs and 14 ALs have been treated in the generation rules. As already commented upon in Section 2.1, the generation rules for the DeCo system fall into four categories, which is illustrated with examples in the following:

(i) Insertion on the left
Example:
“[o],art,def,masc,sing:::”:[s,masc,sing,!def:!def:::]P50;
initial state of the nodelist: *menino*
final state of the nodelist: *o menino*

(ii) Insertion on the right
Examples:
{v,stem,1pes,plural,fut,subj,5,!conjuga:-!conjuga::}:"[armos],dmt,dnp:::"P50;

initial state of the nodelist: *cant*
final state of the nodelist: *cantarmos*

{v,vtd,ação,>obj,npred:->obj,+>od::}:"s,<obj:-<obj,+nod:obj:"P100;
initial state of the nodelist: constr
final state of the nodelist: constr torre

(iii) Alteration of attributes
Example:
{suj,masc:::}:{adj,psuj,!concorda(gen):-!concorda(gen),+masc::}P175;
initial state of the nodelist: *menino bonit*
final state of the nodelist: *menino bonit(>masc)*

(iv) Backtracking
Example:
{:::}?{plural(alomorfe),@pl:::}P250;
initial state of nodelist: *intenção*
final state of nodelist: *intençõ*

The rules created so far may generate all clauses from the two first sentences of the corpus, shown in Figure 8. There are 5.376 rules, distributed as follows:
Insertion on the right: 26
Insertion on the left: 5263
Alteration of attributes: 86
Backtracking: 1

It is to be noted that the current rules generate only simple clauses (with no subordinated or coordinated clauses). The large number of rules for insertion on the left is due to the generation of inflected forms for the verbs, since there are 5247 rules for generating all verb forms for Portuguese. On the other hand, there is only one backtracking rule which points to non determinism and affects positively the computational performance.

A torre de Babel.
Os homens construíram uma torre imensa.
Os homens há muito tempo construíram uma torre imensa na cidade de Babilônia.
Os homens há muito tempo começaram a construir uma torre imensa na cidade de Babilônia.
A torre alcançou os céus.
A torre imensa alcançou os céus.
Uma torre imensa pareceu alcançar os céus.
Uma torre imensa pareceu quase alcançar os céus.
Deus esteve ciente das intenções dos homens.
Deus ficou irado.
Ele criou uma multiplicidade de línguas.
Eles comunicam um com o outro.
Eles são capazes de comunicar um com o outro.
Eles seriam capazes de comunicar um com o outro.
Eles não mais seriam capazes de comunicar um com o outro.

Figure 8. Set of clauses generated from two sentences of the UNL corpus

5. Conclusions and Further Work

Since April, 1997, we have been addressing the two main tasks requested by the UNL Center. A methodology has been established for ascribing Portuguese headwords to the Universal Words (UWs) which will allow us to cover the most relevant words for Portuguese as well as English within the first year of the project. This means that the lexicon available will not be a limitation for tests to be made with the decodifier using real text. We are also investigating possible ways to automatically interpret restrictions in the UW dictionary, so that the process of identifying Portuguese headwords may be speeded up considerably in the near future. As for the creation of generation rules to feed in the DeCo system, our target is to cover the clauses of the sentences in the UNL corpus, for which a UNL representation is available. Though targeted at a small number of sentences, the system developed comprises rules that are sufficiently generic for permitting the generation of a significant number of clauses for Brazilian Portuguese.

6. References

[1] Universal Networking Language: An Electronic Language for Communication, Understanding and Collaboration. Booklet from UNL Center, IAS/UNU, Tokyo, 1996.

[2] Longman Dictionary of Contemporary English, 3ª. ed., 1995.

[3] “As manifestações morfossintáticas da linguagem UNL no Português do Brasil”. Technical Report NILC-97-1, Notas do ICMSC-USP, São Carlos, October, 1997 (in press).