3rd UNITEX/GramLab Workshop
Tours, October, 2014

Révision e adaptation des dictionnaires et graphes d'Unitex-PB à la nouvelle orthographe du portugais

Natalia Perussi CALCIA (PPGL/UFSCar-CNPq)
Adriele Beatriz KUCINSKAS (UFSCar-CNPq)
Marcelo MUNIZ (Dicionário Informal)
Maria das Graças Volpe NUNES (NILC-USP)
Oto Araujo VALE (NILC-UFSCAR/ CENTAL-UCLouvain/ Fapesp)

Unitex-PB

- Ressources du Portugais du Brésil
- Muniz et al (2005) à partir des dictionnaires du projet ReGra (Nunes et al, 1999) et du dictionnaire de conjugaison verbale de Vale (1989)

Unitex-PB

- - DELAS_PB (~67.500 entrées)
- - DELAF_PB (~880.000 entrées)
- - DELACF_PB (~4.000 entrées)

Unitex-PB

- 107 modèles de conjugaison verbale
- 392 modèles de flexion nominale
- 242 modèles de flexion adjectivale

Nouvelle Orthographe du portugais

- “Acordo Ortográfico de 1990” signé par tous les pays lusophones (Angola, Brésil, Cap Vert, Guinée Bissau, Mozambique, Portugal, St Tomé et Principe, Timor Est)

- Simplification et unification de l’orthographe

Simplification et unification de l'ortographe

- Réintroduction des lettres K, W et Y dans l'alphabet
- Suppression du trema
    - **lingüiça** => **linguiça** (Brésil)
- Suppression des consonnes muettes
    - **acção** => **ação** (Portugal)
- Suppression de certains accents:
    - **abençôo** => **abençoo** (Brésil)
    - **idéia** => **ideia** (Brésil)

Unification de l’orthographe

- Règles pour l’utilisation du trait d’union et certains prefixes:
    - **auto-escola** => **autoescola**
    - **contra-senso** => **contrassenso**
    - **extra-conjugal** => **extraconjugal**
    - **semi-árido** => **semiárido**
- MAIS:
    - **auto-observação**
    - **contra-ataque**
    - **semi-interno**

Unification de l’orthographe

- Règles pour l’utilisation du trait d’union:
  - **bico de papagaio** (bec de peroquet)
  - **bico de papagaio** (Ostéophyte)
  - **bico-de-papagaio** (Poinsettia)

Actualisation du dictionnaire

- Adéquation des graphes de flexion nominales et adjectivales

- Vérification des listes des entrées du DELAS e DELAF

- Construction des graphes de conjugaison verbale

Actualisation des graphes

- Flexion nominale:
    - 10 graphes modifiés
- Flexion adjectivale:
    - 6 graphes modifiés
- Conjugaison verbale:
    - au moins 12 graphes supprimés

Actualisation du dictionnaire

- Comparaison automatique avec des listes utilisées par des logiciels propriétaires
- sur les 880.000 formes, 1.287 ont été modifiés
- introduction de 7.900 nouvelles entrés

Graphes de conjugaison verbale

- Muniz (2004) avait adapté directement le fichier des conjugaisons por générer les formes fléchies

- Nécessité de construire des graphes qui décrivent tous les phénomènes de la conjugaison portugaise

enclise et mesoclise

- Conjugaison avec les clitiques:
    - **Ana vem buscar-me** (Ana vient me chercher)
    - **Ana vem buscá-lo** (Ana vient le chercher)
- Mésoclise:
    - **Buscar-te-ei** (je te chercherai)
    - **Buscá-lo-ei** (je le chercherai)

DLF: 15244 simple-word lexical entries

```
a,.ABREV:ms
a,.N:ms
a,.PREP
à,ao.PREPXDET+Art+Def:fs
à,ao.PREPXPRO+Dem:fs
a,ele.PRO+Pes:A3fs
a,o.DET+Art+Def:fs
a,o.PRO+Dem:fs
aba,.N:fs
aba,abar.V:P3s:Y2s
abacaxi,.A:ms
```

DLC: 32 compound lexical entries

```
alto-relevo,.N+AN:ms
amor-perfeito,.N+NA:ms
amor-próprio,.N+NA:ms
azul-celeste,.A+NA:ms:mp:fs:
beija-flor,.N+VN:ms
bem-aventurados,bem-aventura
bem-aventurados,bem-aventura
bem-aventurança,.N+ADVN:fs
bem-feito,.A+ADVA:ms
bem-feito,bem-fazer.V:K
bem-sucedido,.A+ADVA:ms
boa-noite,.N+AN:ms
```

ERR: 385 unknown simple words

Filter unknown words with tags.ind

```
abandoná
abatê
aborrecê
abstemo
acabrunhá
achá
acompanhá
acreditá
ademanes
admirá
afogá
afoutamente
afoutezas
afoutos
afrontá
afundá
aguerrir
ajudá
Alcazar
alhambre
alheá
alicantinas
almeia
amá
```

Trois solutions possibles

- Considérer **buscá** comme une forme verbale
- Utilisation du mode morphologique
- Intégrer les formes enclitiques e mésoclitiques au dictionnaire

Considerer **buscá** comme une forme verbale

```
busca,.N:fs
busca,buscar.V:P2s:P4s:P3s:Y2s
buscá,buscar.V:R:W:V1s:V2s:V4s:V3s
buscam,buscar.V:P4p:P3p
buscar,.V:R:U1s:U4s:U3s:W:V1s:V4s:V3
buscar,buscará.V:F4s:F3s
buscar,buscarão.V:F4p:F3p
buscar,buscarás.V:F2s
buscar,buscarei.V:F1s
buscar,buscareis.V:F2p
buscar,buscaremos.V:F1p
buscar,buscaria.V:C1s:C4s:C3s
buscar,buscariam.V:C4p:C3p
buscar,buscaríamos.V:C1p
buscar,buscarias.V:C2s
buscar,buscaríeis.V:C2p
```

Considerer **buscá** comme une forme verbale

- Problèmes:
  - **buscá** n'est pas une forme verbale indépendante, mais existe seulement en fonction de l'introduction de certains pronoms en position d'enclise ou de mésoclise
  - on peut confondre **buscá** avec de transcription des formes 'populaires': la prononciation 'non soutenue' de la forme infinitive **[bus'ka]**

Utilisation du mode morphologique

- Problèmes:
  - à chaque texte analysé il faudrait faire un pas supplémentaire
  - les formes enclitiques et mésoclitiques devraient êtres traitées séparément

Intégrer les formes enclitiques e mésoclitiques au dictionnaire

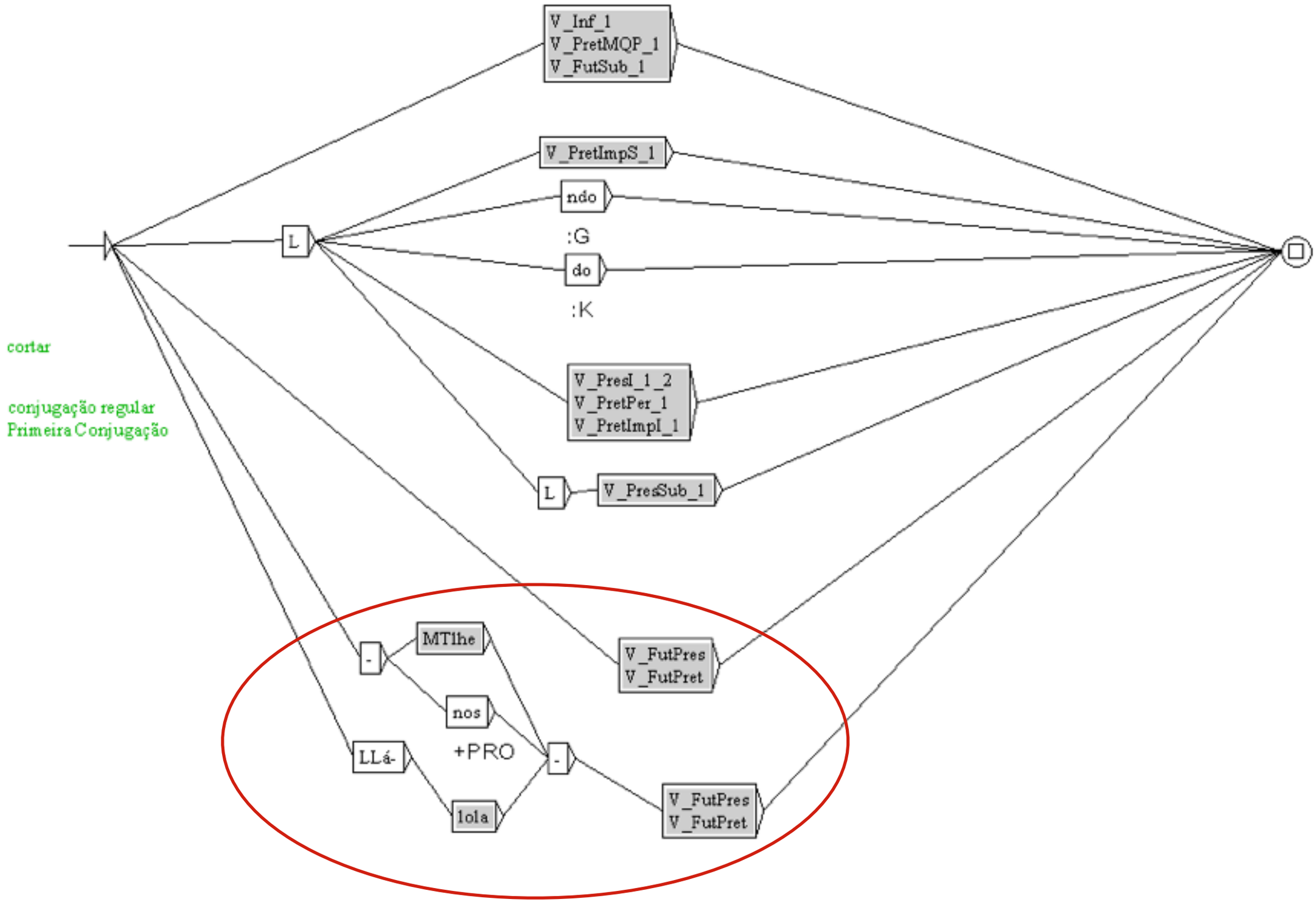
- Chaque forme verbe+clitique sera considérée comme une forme composée

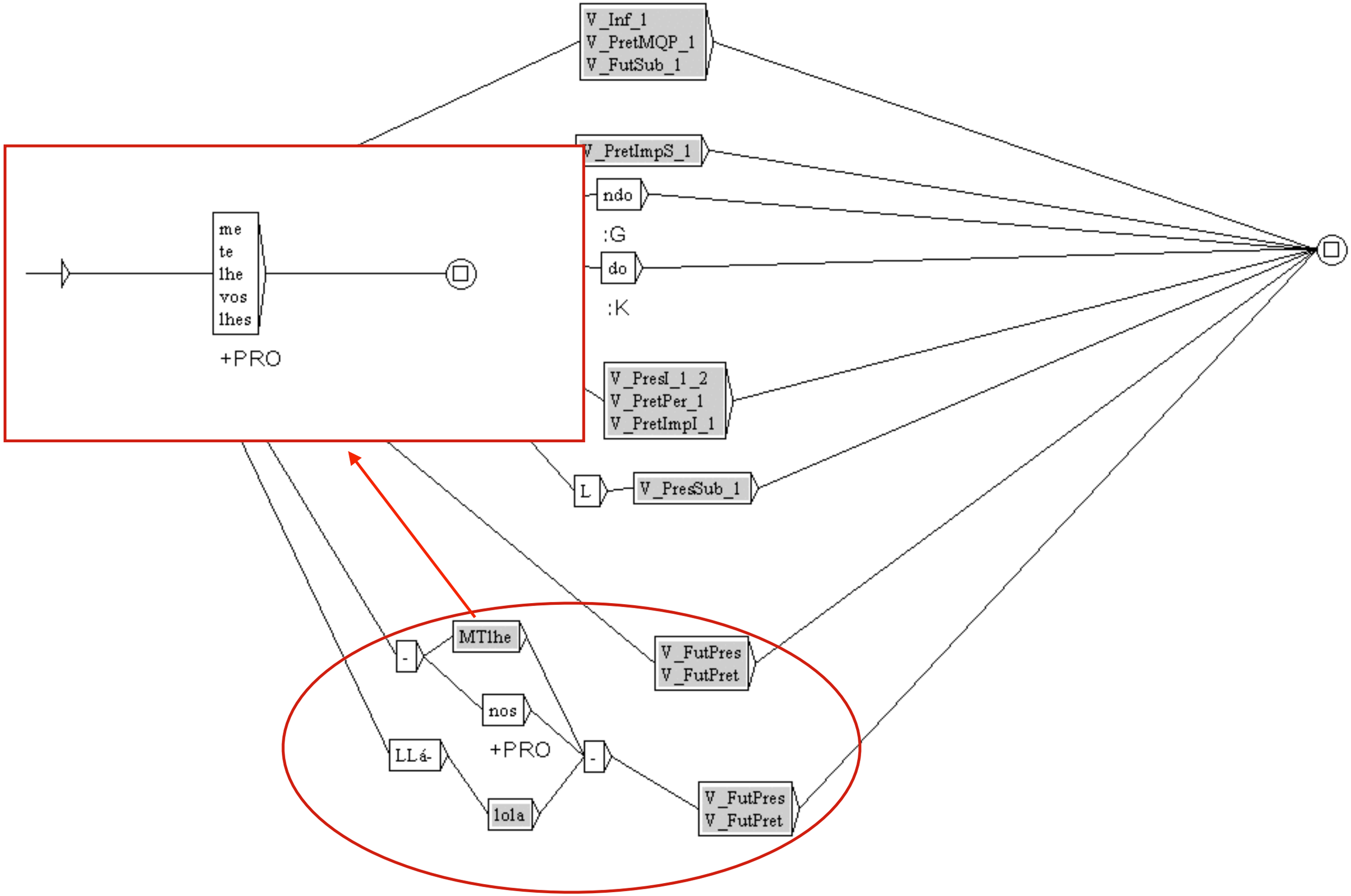
- Nécessité d'une description plus détaillée des clitiques

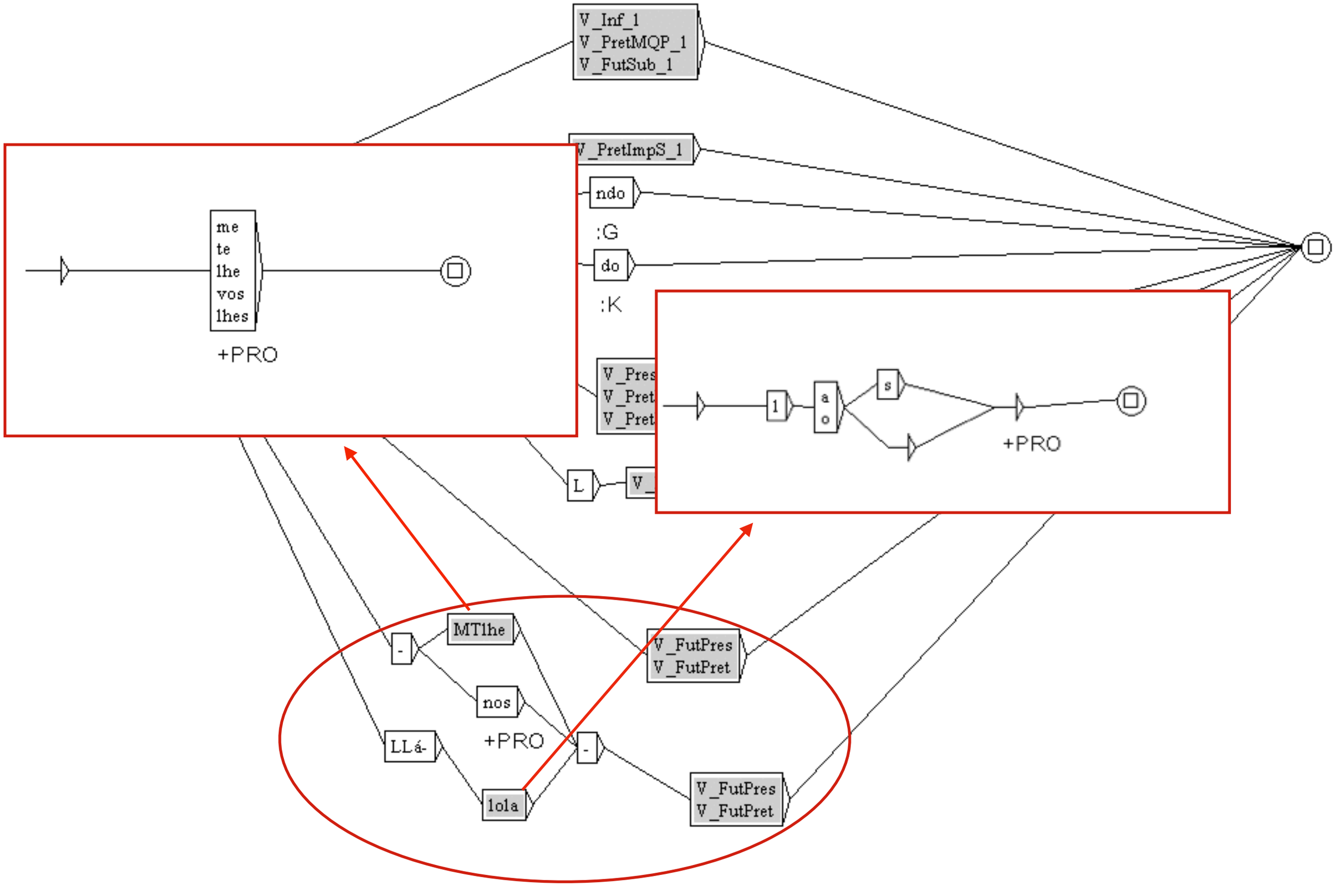
- Problème:

    - explosion du nombre des formes (> 10.000.000 d'entrées au DELAF)

Intégrer les formes mésoclitiques au dictionnaire




V_Inf_1
V_PretMQP_1
V_FutSub_1
V_PretImpS_1
ndo
:G
do
:K
me
te
lhe
vos
lhes
+PRO
V_PresI_1_2
V_PretPer_1
V_PretImpI_1
L
V_PresSub_1
MTlhe
-
nos
+PRO
LLá-
lola
-
V_FutPres
V_FutPret
V_FutPres
V_FutPret


V_Inf_1
V_PretMQP_1
V_FutSub_1
V_PretImpS_1
ndo
:G
do
:K
me
te
lhe
vos
lhes
+PRO
V_Pres
V_Pret
V_Pret
L
V_
1
a
o
s
+PRO
MT1he
-
nos
+PRO
LLá-
1o1a
-
V_FutPres
V_FutPret
V_FutPres
V_FutPret

```
corte,cortar.V:S1s:S3s:Y3s
cortei,cortar.V:J1s
corteis,cortar.V:S2p
cortem,cortar.V:S3p:Y3p
cortemos,cortar.V:S1p:Y1p
cortes,cortar.V:S2s
corto,cortar.V:P1s
cortou,cortar.V:J3s
cortá-la,cortar.V+PRO:U1s:U3s:W1s:W3s
cortá-la-ei,cortar.V+PRO:F1s
cortá-la-eis,cortar.V+PRO:F2p
cortá-la-emos,cortar.V+PRO:F1p
cortá-la-ia,cortar.V+PRO:C1s:C3s
cortá-la-iam,cortar.V+PRO:C3p
cortá-la-ias,cortar.V+PRO:C2s
cortá-la-á,cortar.V+PRO:F3s
cortá-la-ás,cortar.V+PRO:F2s
cortá-la-ão,cortar.V+PRO:F3p
cortá-la-íamos,cortar.V+PRO:C1p
cortá-la-íeis,cortar.V+PRO:C2p
cortá-las,cortar.V+PRO:U1s:U3s:W1s:W3s
cortá-las-ei,cortar.V+PRO:F1s
cortá-las-eis,cortar.V+PRO:F2p
cortá-las-emos,cortar.V+PRO:F1p
cortá-las-ia,cortar.V+PRO:C1s:C3s
cortá-las-iam,cortar.V+PRO:C3p
cortá-las-ias,cortar.V+PRO:C2s
cortá-las-á,cortar.V+PRO:F3s
cortá-las-ás,cortar.V+PRO:F2s
cortá-las-ão,cortar.V+PRO:F3p
cortá-las-íamos,cortar.V+PRO:C1p
cortá-las-íeis,cortar.V+PRO:C2p
cortá-lo,cortar.V+PRO:U1s:U3s:W1s:W3s
cortá-lo-ei,cortar.V+PRO:F1s
```

Prochains pas

- Réviser, encore, les graphes de flexion nominale et adjectivale
- Etiqueter les pronoms des enclises et mesoclises
- Traitement des formes composées

Références

- Muniz, M. C.M., Nunes, M. G. V and Laporte, E. (2005) UNITEX-PB, a set of flexible language resources for Brazilian Portuguese. Workshop on Technology on Information and Human Language (TIL), 2005, São Leopoldo, Brazil. pp. 2059-2068

- Nunes, M. G. V., F. M. C. Vieira, C. Zavaglia, C. R. C. Sossolote, & J. Her- nandez (1996). A construção de um léxico de português do brasil: Lições aprendidas e perspectivas. In Anais do II Workshop de Processamento Computacional de Português Escrito e Falado (PROPOR´96), pp. 61–70. CEFET-PR, Curitiba.

- Ranchhod, E., Mota, C., and Baptista, J. (1999). A computational lexicon of Portuguese for automatic text parsing. In Proceedings of SIGLEX99: Standardizing Lexical Resources, 37th Annual Meeting of the ACL, pages 74–80. College Park, Maryland, USA.

- Vale, O.V. Dictionnaire électronique des conjugaisons des verbes du portugais du Brésil. Rapport Technique du LADL n 27, Paris : Université Paris 7. 1990.