



**INTERNATIONAL NOOJ 2020
CONFERENCE**

JUNE, 05-07, 2020. Zagreb , Croatia

Standardization and Implementation of Lexicon-Grammar Tables in NooJ Platform

Asmaa Kourtin, Mohammed Mourchid, Abdelaziz
Mouloudi and Samir Mbarki

MISC Laboratory, Faculty of Science,

Ibn Tofail University, Kenitra-Morocco

Outline

- Introduction
- Overview of the LG approach & previous work
- LG tables properties standardization
- French LG tables standardization
- Implementation
- Conclusion & perspectives

Introduction

- LG approach is very important in automatic NLP.
- Consists of studying the language lexicon.
- French LG tables constitute a large lexical, syntactic and semantic linguistic resources.
- They couldn't been used directly in NooJ as they are.
- The representation format of their properties is not compatible with NooJ properties.

Introduction

Objective :

- Standardize all the properties used in French LG tables.
- Standardize French LG tables.
- Implement the standardized tables in NooJ plateform :
 - Generate NooJ dictionaries.
 - Create syntactic grammars.

The Lexicon-Grammar approach

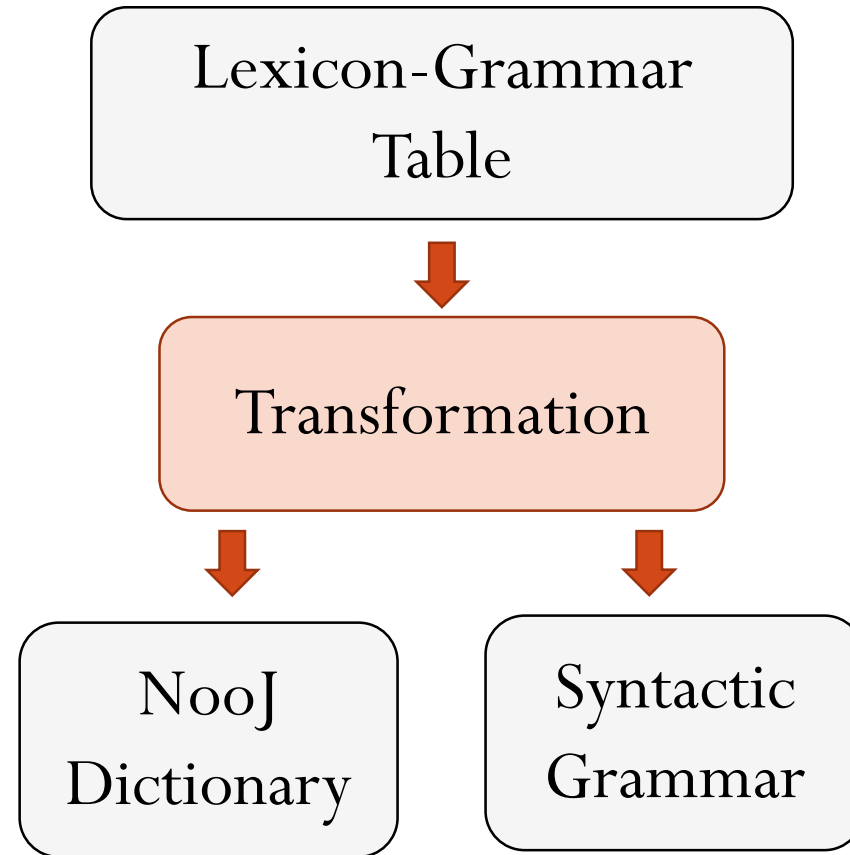
- Initiated by Gross and his LADL's team in 1975.
- A Lexicon-Grammar table groups a number of entries sharing the same definitional construction.

| <ID> | N0 =: Nhum | N0 =: N-hum | Ppv =: se figé | Ppv =: en figé | Ppv =: y figé | Ppv =: Neg | <ENT>Ppv | <ENT>V | N0 V | <ENT>Det1 | N0 V Det1 C1 Prép2 N2 | <ENT>C1 | C1 =: Npc | [passif] | Exemple |
|------|------------|-------------|----------------|----------------|---------------|------------|----------|------------|------|-----------|-----------------------|-------------------|-----------|----------|----------------------------------|
| 1 | + | - | - | - | - | - | <E> | abandonner | - | la | - | compétition | - | + | §abandonner§la compétition |
| 2 | + | - | - | - | - | - | <E> | abandonner | - | le | - | domicile conjugal | - | + | §abandonner§le domicile conjugal |
| 3 | + | - | - | - | - | - | <E> | abandonner | - | les | - | lieux | - | + | §abandonner§les lieux |
| 4 | + | - | - | - | - | - | <E> | abandonner | + | la | - | partie | - | + | §abandonner§la partie |
| 5 | + | - | - | - | - | - | <E> | abandonner | - | la | + | pose | - | + | §abandonner§la pose |
| 6 | + | - | - | - | - | - | <E> | abandonner | - | le | - | terrain | - | + | §abandonner§le terrain |
| 7 | + | - | - | - | - | - | <E> | abjurer | - | la | - | foi catholique | - | + | §abjurer§la foi catholique |
| 8 | - | + | - | - | - | - | <E> | abolir | - | le | - | temps | - | + | §abolir§le temps |
| 9 | + | + | - | - | - | - | <E> | accélérer | - | le | - | mouvement | - | + | §accélérer§le mouvement |

Excerpt of the lexicon-grammar table C1D for French frozen expressions

The Lexicon-Grammar approach :

Integration in NooJ (Silberztein, 2015)



The Lexicon-Grammar approach : Integration in NooJ

- The automation advantages :
 - Optimize time.
 - Optimize human effort.

 Dictionaries automatic generation
(Kourtin et al., 2020)

The Lexicon-Grammar approach : Integration in NooJ

Motivations :

- The automatic generation of NooJ dictionaries suffered from the non-standardization of all the potential Lexicon-Grammar tables :
 - The representation format of syntactico-semantic properties is not compatible with NooJ.
 - There is different names for the same property.
 - There is different properties for the same name.



Non-existence of a standard helping linguists to define unified properties names.

The Lexicon-Grammar approach : Integration in NooJ

Motivations :

The non-standardization of all the potential lexicon-grammar tables

| | | |
|---------------------|----------|---|
| N0= : "Inr | | |
| N0= : "le fait Qu P | | |
| N0= : "V1-Ω | | |
| verbe | | |
| V concret | | |
| N0 V | | |
| Adjectif = ant | | |
| Adjectif = able | | |
| Adjectif = eux | | |
| Adjectif =(E+at)eur | | |
| N1= : " Nhum | | |
| N1= : "-Nhum | | |
| Le fait Qu P | | |
| N1 se V de ce que P | | |
| ... | | |
| + | calmer | + |
| + | captiver | . |

Excerpt of a table of Maurice Gross

Excerpt of a table of simonetta Vietri

| | | | | | | | | | | | | | | | | | |
|------|-----------|-----------|----------|----------------|----------------|------------|---------|-----|--------------|-------------|--------------|------|-------|-----------------|------------------|---------------|-----|
| <ID> | N0 = Nhum | N0 =: Nnc | <ENT>Ppv | Ppv =: se figé | Ppv =: en figé | Ppv =: Neg | <ENT>V | Neg | Aux =: avoir | Aux =: être | N0 est Vpp W | N0 V | Prép1 | N0 V Prép N1hum | N0 V Prép N1-hum | Prép N1 = Ppv | ... |
| 1 | + | . | <E> | . | . | . | achever | . | + | . | . | . | de | . | . | . | . |
| 2 | + | + | <E> | . | . | . | aller | . | . | . | . | . | <E> | . | . | . | . |

Excerpt of a table of Max Silberztein

| | | | | | | | | | | | | | | | | | |
|---|------------------------------|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|--|
| + | [1] N0 = +hum | | | | | | | | | | | | | | | | |
| . | [2] N0 = -hum | | | | | | | | | | | | | | | | |
| . | [3] N0 = ChF | | | | | | | | | | | | | | | | |
| . | Neg | | | | | | | | | | | | | | | | |
| . | Pro | | | | | | | | | | | | | | | | |
| . | allungare | | | | | | | | | | | | | | | | |
| . | alzare | | | | | | | | | | | | | | | | |
| . | non | | | | | | | | | | | | | | | | |
| . | il | | | | | | | | | | | | | | | | |
| . | un | | | | | | | | | | | | | | | | |
| . | muso | | | | | | | | | | | | | | | | |
| . | dito | | | | | | | | | | | | | | | | |
| + | [4] C1 = body-part | | | | | | | | | | | | | | | | |
| . | [5] C1 = plural | | | | | | | | | | | | | | | | |
| . | [6] Unaccusative DET C1 si V | | | | | | | | | | | | | | | | |
| . | ... | | | | | | | | | | | | | | | | |

LG tables properties standardization

- French LG tables constitute a very rich linguistic resources.
- List the properties used in the French LG tables.
 - 515 properties for verbs tables
 - 225 properties for frozen expressions
 - ...
- Propose an unified meaningful name for each property to be compatible with the NooJ properties format.

LG tables properties standardization

| Property name before standardization | Property name after standardization | Meaning | Number of occurrences | Tables |
|--------------------------------------|-------------------------------------|--|-----------------------|--|
| NHum N =: Nhum | H, H0, H1, H2, ... | Human or animated noun | 146 | V_1, V_3-V_16, V_18, V_31R, V_32H, V_32R3, V_33, V_34L0, V_35L, V_35LD, ... |
| N-Hum N =: N-Hum | NH, NH0, NH1, NH2, ... | Non-human noun | 96 | V_1, V_3, V_5-V_16, V_18, V_32A, V_32CL, V_32CV, V_32D, V_32NM, ... |
| N0 =: Nabs NAbstrait | A, A0, A1, A2, ... | Abstract noun | 10 | V_4, V_31H, V_32C, V_32D, V_32H, V_35L, V_35LD, V_35LR, V_35LS, V_36DT |
| N0 =: Nnr | Nr, Nr0, Nr1, Nr2, ... | Noun denoting a person, a concrete object, an abstract entity, a completion or an infinitive | 26 | V_3, V_6-V_16, V_18, V_31R, V_32H, V_32NM, V_32R2, V_33, V_34L0, V_35L, V_35LD, V_35LR, V_35LS, V_35R, V_35S, V_35ST |
| | | | | |

LG tables properties standardization

Some LG properties standardization conditions :

- Must have a meaningful name that will be simply used in NooJ.
- Must not contain symbols not recognized by dictionaries or special characters (such as +, #, ...).

French LG tables standardization

- Use the proposed unified properties to standardize the LG tables.
- C1D table for French frozen expressions
 - Basic construction : N0 V Det1 C1

| <ID> | N0 =: Nhum | N0 =: N-hum | Ppv =: se figé | Ppv =: en figé | Ppv =: y figé | Ppv =: Neg | <ENT>Ppv | <ENT>V | N0 V | <ENT>Det1 | N0 V Det1 C1 Prép2 N2 | <ENT>C1 | C1 =: Npc | [passif] | Exemple |
|------|------------|-------------|----------------|----------------|---------------|------------|----------|------------|------|-----------|-----------------------|-------------------|-----------|----------|----------------------------------|
| 1 | + | - | - | - | - | - | <E> | abandonner | - | la | - | compétition | - | + | §abandonner§la compétition |
| 2 | + | - | - | - | - | - | <E> | abandonner | - | le | - | domicile conjugal | - | + | §abandonner§le domicile conjugal |
| 3 | + | - | - | - | - | - | <E> | abandonner | - | les | - | lieux | - | + | §abandonner§les lieux |
| 4 | + | - | - | - | - | - | <E> | abandonner | + | la | - | partie | - | + | §abandonner§la partie |
| 5 | + | - | - | - | - | - | <E> | abandonner | - | la | + | pose | - | + | §abandonner§la pose |
| 6 | + | - | - | - | - | - | <E> | abandonner | - | le | - | terrain | - | + | §abandonner§le terrain |
| 7 | + | - | - | - | - | - | <E> | abjurer | - | la | - | foi catholique | - | + | §abjurer§la foi catholique |
| 8 | - | + | - | - | - | - | <E> | abolir | - | le | - | temps | - | + | §abolir§le temps |
| 9 | + | + | - | - | - | - | <E> | accélérer | - | le | - | mouvement | - | + | §accélérer§le mouvement |

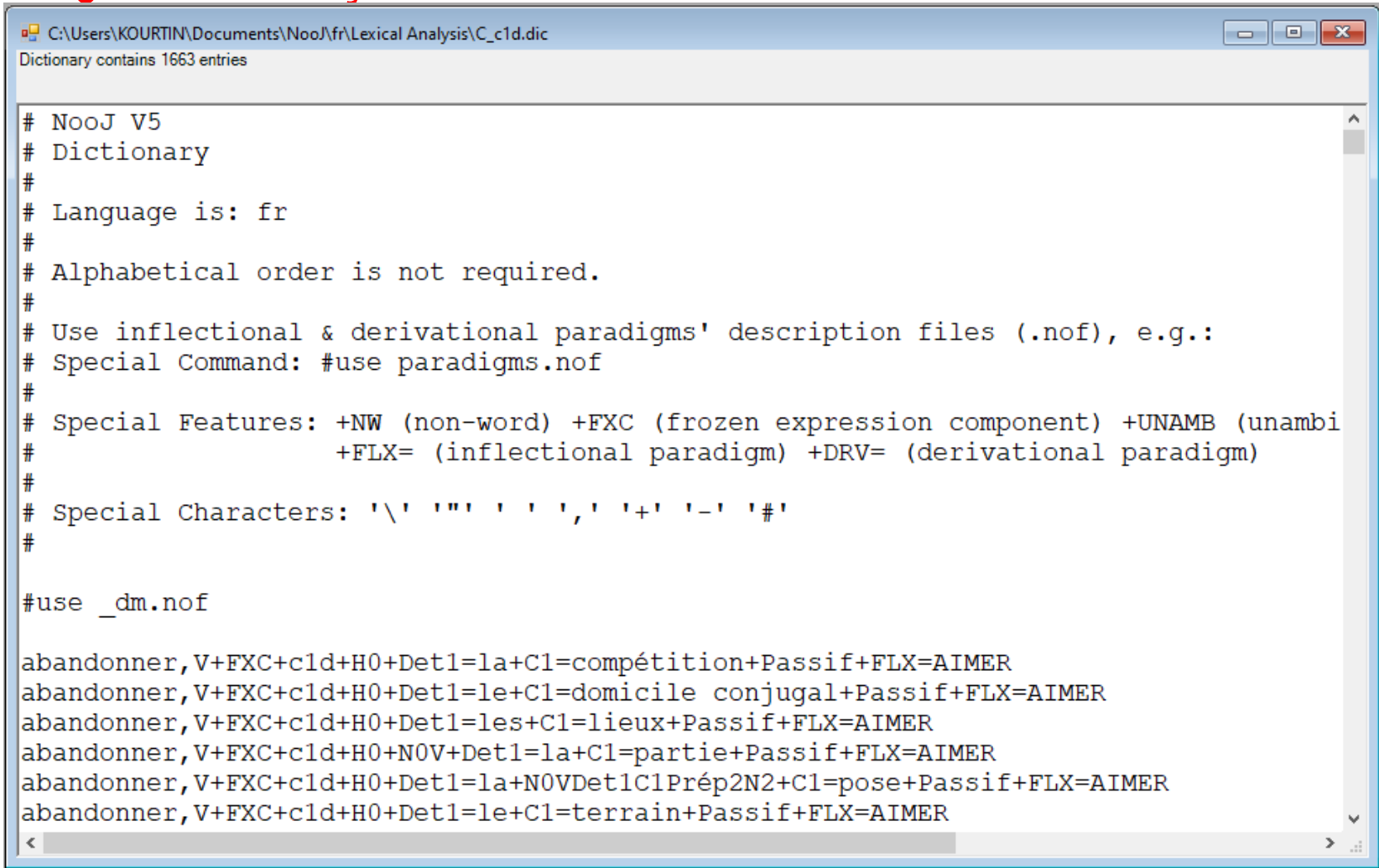
French LG tables standardization

| <ID> | H0 | NH0 | Ppv=SeFigé | Ppv=EnFigé | Ppv=YFigé | Ppv=Neg | Ppv | V | NOV | Det1 | NOVDet1C1Prép2N2 | C1 | C1pc | Passif | <OPT>Exemple |
|------|----|-----|------------|------------|-----------|---------|-----|------------|-----|------|------------------|-------------------|------|--------|----------------------------------|
| 1 | + | - | - | - | - | - | <E> | abandonner | - | la | - | compétition | - | + | §abandonner§la compétition |
| 2 | + | - | - | - | - | - | <E> | abandonner | - | le | - | domicile conjugal | - | + | §abandonner§le domicile conjugal |
| 3 | + | - | - | - | - | - | <E> | abandonner | - | les | - | lieux | - | + | §abandonner§les lieux |
| 4 | + | - | - | - | - | - | <E> | abandonner | + | la | - | partie | - | + | §abandonner§la partie |
| 5 | + | - | - | - | - | - | <E> | abandonner | - | la | + | pose | - | + | §abandonner§la pose |
| 6 | + | - | - | - | - | - | <E> | abandonner | - | le | - | terrain | - | + | §abandonner§le terrain |
| 7 | + | - | - | - | - | - | <E> | abjurer | - | la | - | foi catholique | - | + | §abjurer§la foi catholique |
| 8 | - | + | - | - | - | - | <E> | abolir | - | le | - | temps | - | + | §abolir§le temps |
| 9 | + | + | - | - | - | - | <E> | accélérer | - | le | - | mouvement | - | + | §accélérer§le mouvement |

Excerpt of the lexicon-grammar table C1D **after standardization**.

Implementation

NooJ dictionary for C1D table :



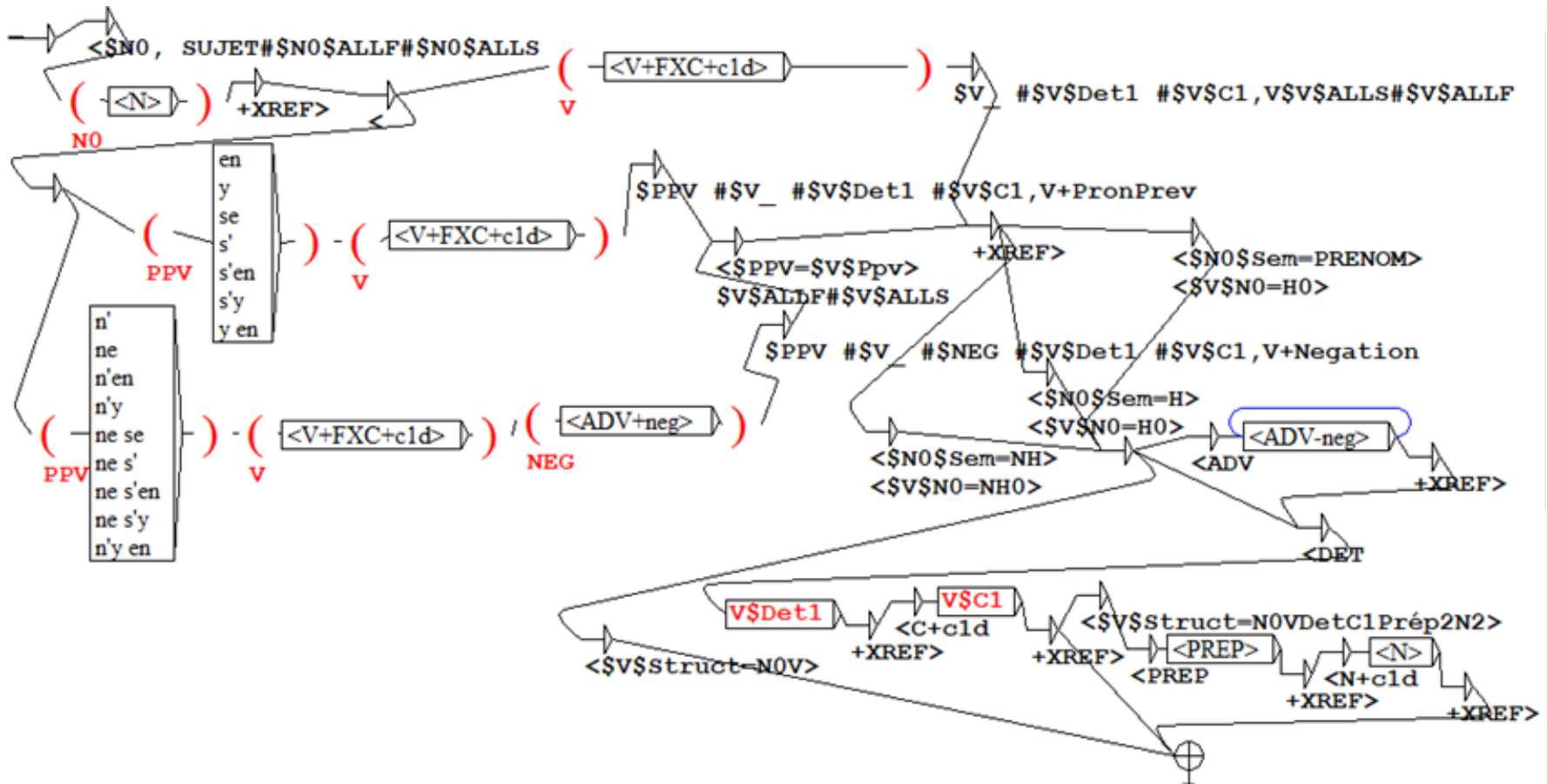
```
C:\Users\KOURTIN\Documents\NooJ\fr\Lexical Analysis\C_c1d.dic
Dictionary contains 1663 entries

# NooJ V5
# Dictionary
#
# Language is: fr
#
# Alphabetical order is not required.
#
# Use inflectional & derivational paradigms' description files (.nof), e.g.:
# Special Command: #use paradigms.nof
#
# Special Features: +NW (non-word) +FXC (frozen expression component) +UNAMB (unambi
#                   +FLX= (inflectional paradigm) +DRV= (derivational paradigm)
#
# Special Characters: '\ ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' '
#
#use _dm.nof

abandonner,V+FXC+cld+H0+Det1=la+C1=compétition+Passif+FLX=AIMER
abandonner,V+FXC+cld+H0+Det1=le+C1=domicile conjugal+Passif+FLX=AIMER
abandonner,V+FXC+cld+H0+Det1=les+C1=lieux+Passif+FLX=AIMER
abandonner,V+FXC+cld+H0+N0V+Det1=la+C1=partie+Passif+FLX=AIMER
abandonner,V+FXC+cld+H0+Det1=la+N0VDet1C1Prép2N2+C1=pose+Passif+FLX=AIMER
abandonner,V+FXC+cld+H0+Det1=le+C1=terrain+Passif+FLX=AIMER
```

Implementation

Syntactic grammar for C1D table : (C_c1d.nog)



C:\Users\KOURTIN\Documents\Noo\fr\Projects\testC_c1d.not

- 11 + / 17 TUs

Characters
Tokens
Digrams

Language is "French(fr)".
Text Delimiter is: \n (NEWLINE)
Text contains 17 Text Units (TUs).
87 tokens including:
85 word forms
2 delimiters

Show Text Annotation Structure

Luc abandonne fréquemment la partie
Luc n'arrange pas fréquemment les choses
Luc ne dépasse pas la dose prescrite

| | | |
|-------------------------------|---|-------------|
| Luc,N+Sem=PRENOM+Genre=m+Nb=s | abandonner la partie,V+c1d+N0=H0+Struct=N0V+Det1=la+C1=partie+Passif+Temps=IP+Pers=2+Nb=s | ADV |
| | abandonner la partie,V+c1d+N0=H0+Struct=N0V+Det1=la+C1=partie+Passif+Temps=IP+Pers=2+Nb=s | ADV |
| | abandonner la partie,V+c1d+N0=H0+Struct=N0V+Det1=la+C1=partie+Passif+Temps=S+Pers=3+Nb=s | ADV |
| | abandonner la partie,V+c1d+N0=H0+Struct=N0V+Det1=la+C1=partie+Passif+Temps=S+Pers=3+Nb=s | ADV |
| | abandonner la partie,V+c1d+N0=H0+Struct=N0V+Det1=la+C1=partie+Passif+Temps=S+Pers=1+Nb=s | ADV |
| | abandonner la partie,V+c1d+N0=H0+Struct=N0V+Det1=la+C1=partie+Passif+Temps=S+Pers=1+Nb=s | fréquent,AI |

C:\Users\KOURTIN\Documents\Noo\fr\Projects\testC_c1d.not

- 11 + / 17 TUs

Characters
Tokens
Digrams

Language is "French(fr)".
Text Delimiter is: \n (NEWLINE)
Text contains 17 Text Units (TUs).
87 tokens including:
85 word forms
2 delimiters

Show Text Annotation Structure

Luc abandonne fréquemment la partie
Luc n'arrange pas fréquemment les choses
Luc ne dépasse pas la dose prescrite

| | | | |
|------|---|---|--------------------------------------|
| lb=s | ADV | DET | C+c1d |
| lb=s | ADV | DET | C+c1d |
| b=s | ADV | DET | C+c1d |
| b=s | ADV | DET | C+c1d |
| b=s | ADV | DET | C+c1d |
| b=s | fréquent,ADV+DRVORIGINALCAT=A+Info=dadj | le,PRO+Dist=clit+Pers=3+Nb=s+Genre=f+Fonc=acc | partir,V+Aux=e+Genre=f+Nb=s+Temps=PP |

Conclusion

- The Lexicon-Grammar tables simplify the uses of the linguistic data and their modifications.
- This standardization :
 - Facilitate more NooJ dictionaries automatic generation from LG tables.
 - Simplify the uses of LG tables in NooJ and allow to profit from their advantages.
 - Have a positive effect on the efficiency of texts and corpora analysis.

Perspectives

- Extend this work to take other languages into account
 - Add new properties.
- Create Lexicon-Grammar tables for the Arabic language.

*Thank
you*

