9. Vocabulary and Morphology

Linguistic resources

- NooJ lexical symbols (*e.g.*, <go+PR+3+s>) and syntactic symbols (*e.g.*, <N+Hum+p>) are defined by linguistic resources.

- These linguistic resources are handcrafted dictionaries and grammars.

- Dictionaries and morphological grammars are typically used to recognize and annotate ALUs

- Syntactic and semantic grammars are used to recognize and annotate more complex linguistic units, *e.g.*, noun phrases, modifiers, sentences.

The vocabulary

- The vocabulary of a language is the set of all its Atomic Linguistic Units (ALUs)

- Because it is no possible to compute the properties of ALUs, we need to list and describe all ALUs *in extenso*

- But is it possible? Is the vocabulary of a natural language finite in size?
    - Languages keep evolving…
    - How many language variants…

Languages keep evolving

Obsolete words:

elflocks: *mangled wavy hair*
crapulous: *to feel ill from excessive eating*
to groak: *to stare at someone who is eating*
to lunt: *to go for a walk and smoke a pipe*
to twattle: *to gossip*
...

New words:
a b-day, to chillax, to cringe, to ghost someone, to mansplain, a photobomb, a staycation...
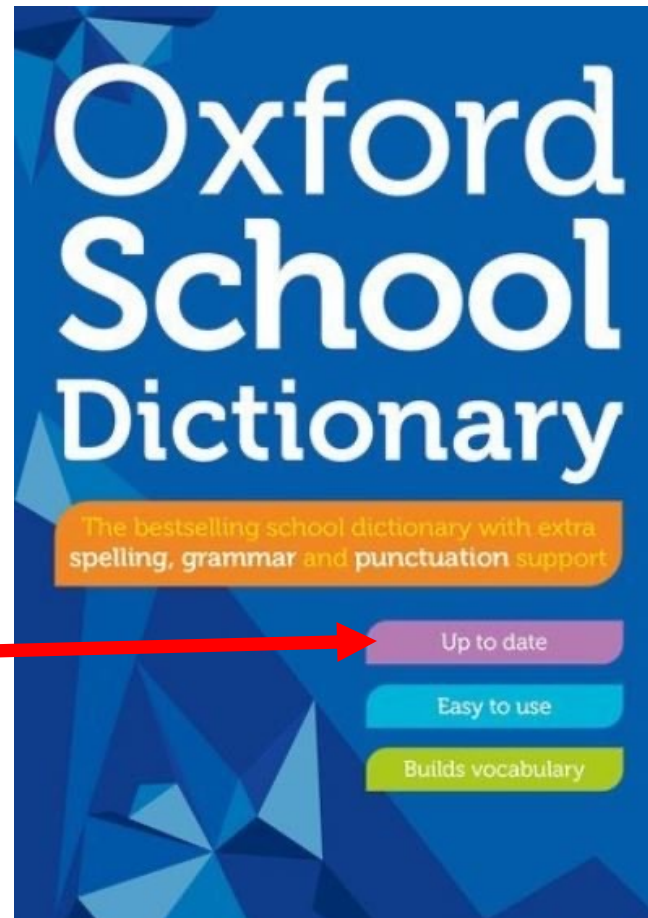

to manspread

First Person Shooter

Smartphones
5-inch and less

Phablets
Between 5-7 inches

Tablets
7-inches and above

Languages keep evolving

Solution: update the lexical resources regularly, *e.g.*, every year

up to date

Many languages variants

Scientific and Technical terms

an activated complex (Chemistry), an object-oriented language (Computer Science), a differencial equation (Mathematics) ...

Arts and Crafts

a boom shot (Movies), a cambré (Ballet), a dominant (Music), an embossing (Paper printing), a parodos (Theater)...

Games, Hobbies, Sports...

a queen gambit (Chess), a double-ace flush (Poker), a direct free kick (American Football), a divot (Golf)...

Many language variants

British English

a bloke, a nitwit, a brolly, a flat, a hoover, the loo...

American English (from Boston, Texas, ...)

a diaper, a faucet, eyeglasses, French fries, hankerin',
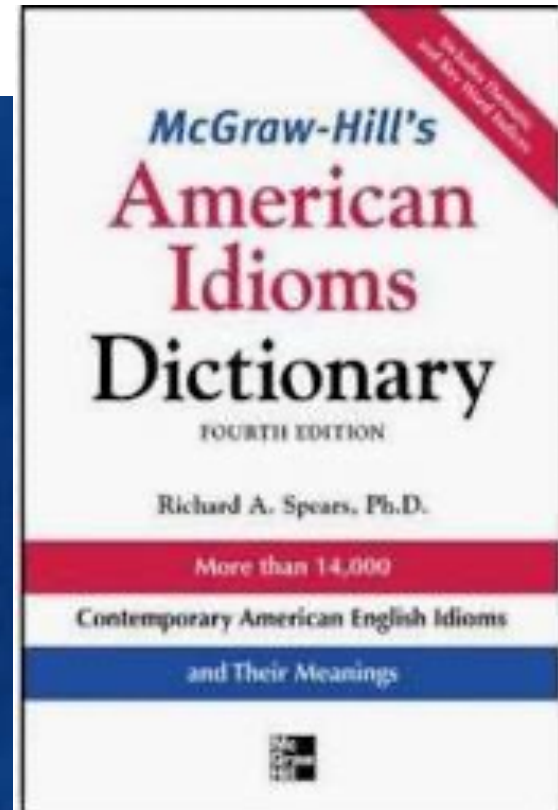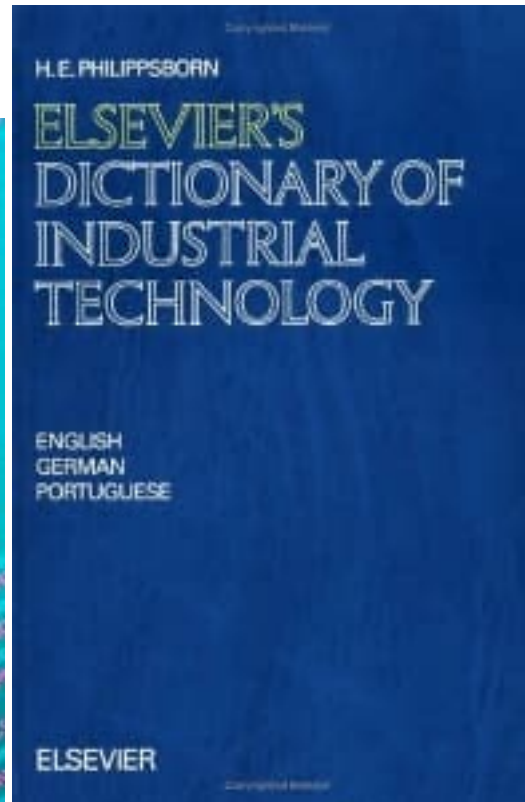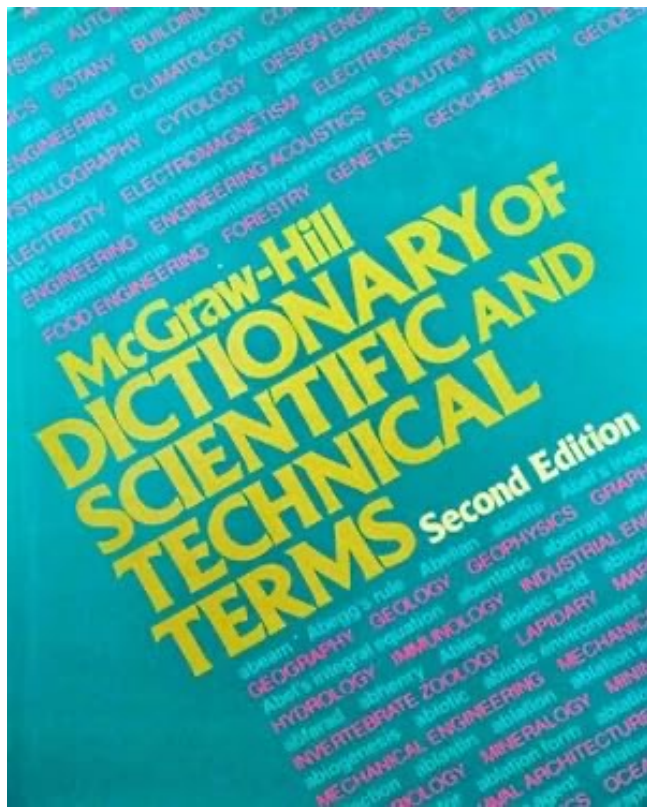a pop, a vacuum cleaner...

Commonwealth English

abi (Nigeria), an arvo (Australia), to bliksem someone
(South Africa), a good name (India), ...

Many language variants

Solution:

- Describe the standard vocabulary: shared by all Native speakers
- Add terminological databases, dictionaries of variants, etc. to the fulfill the needs of a particular application
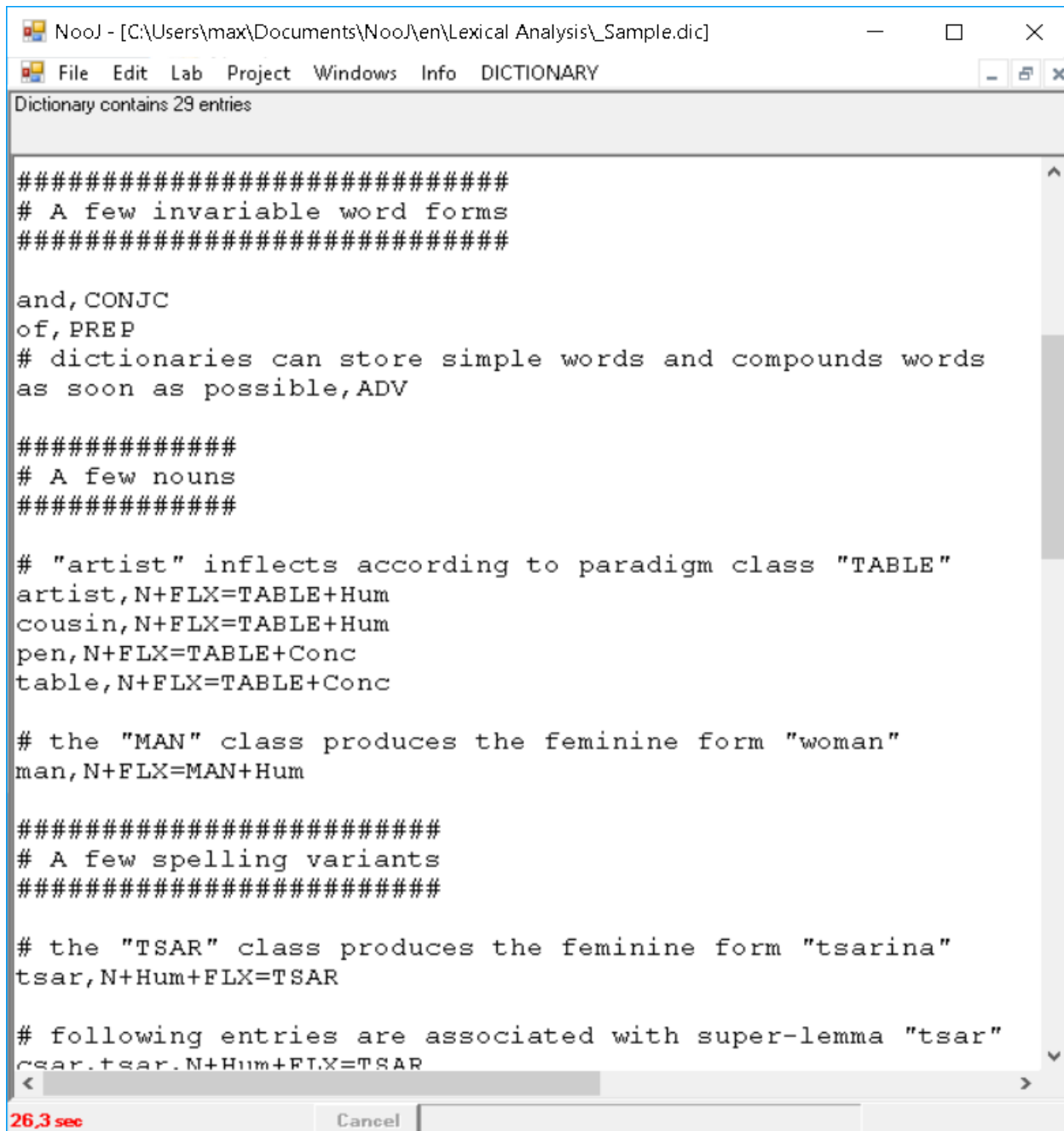
The vocabulary

- The vocabulary is a finite set of elements
- Each element of the vocabulary is atomic, *i.e.*, it cannot be analyzed from its components
- Vocabulary elements are the Atomic Linguistic Units (ALUs)
- Homographs and ambiguous words, (*i.e.*, with different sets of properties) are independent ALUs
- A NLP software application must be able to process four types of ALUs:
  - simple words, *e.g.*, *table*
  - affixes, *e.g.*, *re-*, *-ization*
  - multiword units, *e.g.*, *all of a sudden*, *a blue collar*
  - discontinuous expressions, *e.g.*, *to turn ... off*, *to take ... into account*

NooJ dictionaries

- A NooJ dictionary aims at formalizing the vocabulary precisely:
  - All ALUs must be taken into account
  - Every ALU must be described by one lexical entry
  - All wordforms must be accounted for (*e.g., laughable*)
  - All morphological rules must be explicitly described
  - All rule exceptions must be explicitly listed

NooJ Dictionaries

NooJ - [C:\Users\max\Documents\NooJ\en\Lexical Analysis_Sample.dic]

File   Edit   Lab   Project   Windows   Info   DICTIONARY

Dictionary contains 29 entries

```
###############################
# A few invariable word forms
###############################

and,CONJC
of,PREP
# dictionaries can store simple words and compounds words
as soon as possible,ADV

##############
# A few nouns
##############

# "artist" inflects according to paradigm class "TABLE"
artist,N+FLX=TABLE+Hum
cousin,N+FLX=TABLE+Hum
pen,N+FLX=TABLE+Conc
table,N+FLX=TABLE+Conc

# the "MAN" class produces the feminine form "woman"
man,N+FLX=MAN+Hum

##########################
# A few spelling variants
##########################

# the "TSAR" class produces the feminine form "tsarina"
tsar,N+Hum+FLX=TSAR

# following entries are associated with super-lemma "tsar"
csar,tsar,N+Hum+FLX=TSAR
```

26,3 sec   Cancel

NooJ Dictionaries are associated with inflectional/derivational grammars

```
NooJ - [C:\Users\max\Documer          NooJ - [C:\Users\max\Documents\NooJ\en\Lexical Analysis\_Sample.nof]
File   Edit   Lab   Project   W     File   Edit   Lab   Project   Windows   Info

Dictionary contains 29 entries

####################               TABLE = <E>/s | s/p ;
# A few invariable                 # takes an 's' in plural
####################

                                   CLOTH = <E>/s | es/p;
and,CONJC
of,PREP                            MAN = <E>/m+s | <B2>en/p
# dictionaries can                     | <LW>wo/f+s | <B2>en<LW>wo/f+p;
as soon as possible
                                   TSAR = <E>/m+s | s/p
#############                          | ina/f+s | inas/f+p;
# A few nouns
#############                       ###################################
                                   # Adjective Inflectional Paradigms
# "artist" inflects                ###################################
artist,N+FLX=TABLE+
cousin,N+FLX=TABLE+                A = <E>/n ;
pen,N+FLX=TABLE+Cor
table,N+FLX=TABLE+C                AER = <E>/n | er/cp | est/sp ;

# the "MAN" class                  ############################
man,N+FLX=MAN+Hum                  # Verb Inflection Paradigms
                                   ############################
####################
# A few spelling va                BE = <E>/INF | ing/G | en/PP
####################                    | <BW> (am/PR+s+1 | are/PR+s+2 | is/PR+s+3 | are/PR+p+1 | are/PR+p+2 | are/PR+p+
                                       | <BW> (was/PT+s+1 | were/PT+s+2 | was/PT+s+3 | were/PT+p+1 | were/PT+p+2 | were
# the "TSAR" class
tsar,N+Hum+FLX=TSAR                DO = <E>/INF | ing/G | ne/PP
                                       | <E>/PR+s+1 | <E>/PR+s+2 | es/PR+s+3 | <E>/PR+p+1 | <E>/PR+p+2 | <E>/PR+p+3
# following entries                    | <B> (id/PT+s+1 | id/PT+s+2 | id/PT+s+3 | id/PR+p+1 | id/PR+p+2 | id/PR+p+3);
csar,tsar,N+Hum+FLX

26,3 sec                           26,3 sec                        Cancel
```

NooJ Dictionaries are associated with inflectional/derivational grammars

File Edit Lab Project W

Dictionary contains 29 entries

```
##################
# A few invariable
##################

and,CONJC
of,PREP
# dictionaries can
as soon as possible

#############
# A few nouns
#############

# "artist" inflect
artist,N+FLX=TABLE+
cousin,N+FLX=TABLE+
pen,N+FLX=TABLE+Con
table,N+FLX=TABLE+C

# the "MAN" class
man,N+FLX=MAN+Hum

##################
# A few spelling va
##################

# the "TSAR" class
tsar,N+Hum+FLX=TSAR

# following entries
csar.tsar.N+Hum+FLX
```

NooJ - [C:\Users\max\Documents\NooJ\en\Lexical Analysis_Sample.nof]

File Edit Lab Project Windows Info

```
TABLE = <E>/s | s/p ;
# takes an 's' in plural

CLOTH = <E>/s | es/p;

MAN = <E>/m+s | <B2>en/p
      <LW>wo/f+s | <B2>en<LW>wo/f+p;

TSAR = <E>/m+s | s/p
       | ina/f+s | inas/f+p;

##################################
# Adjective Inflectional Paradigms
##################################

A = <E>/n ;

AER = <E>/n | er/cp | est/sp ;

###########################
# Verb Inflection Paradigms
###########################

BE = <E>/INF | ing/G | en/PP
     | <BW> (am/PR+s+1 | are/PR+s+2 | is/PR+s+3 | are/PR+p+1 | are/PR+p+2 | are/PR+p+
     | <BW> (was/PT+s+1 | were/PT+s+2 | was/PT+s+3 | were/PT+p+1 | were/PT+p+2 | were

DO = <E>/INF | ing/G | ne/PP
     | <E>/PR+s+1 | <E>/PR+s+2 | es/PR+s+3 | <E>/PR+p+1 | <E>/PR+p+2 | <E>/PR+p+3
     | <B> (id/PT+s+1 | id/PT+s+2 | id/PT+s+3 | id/PR+p+1 | id/PR+p+2 | id/PR+p+3);
```

Croatian Dictionary (Kristina Kocijan, Univ. of Zagreb)

```
File   Edit   Lab   Project   Windows   Info        DICTIONARY
Dictionary contains 20136 entries


#use hrCroNoun.nof

#IMENICE

fakultet,N+c+m+FLX=ALAT

abdiciranje,N+c+n+FLX=CRTANJE
abdikacija,N+c+f+FLX=MEDO
abdukcija,N+c+f+Domena=MED+DomenaType=PROC+FLX=MEDO
abeceda,N+c+f+FLX=MEDO
aberacija,N+c+f+FLX=MEDO
abolicija,N+c+f+FLX=MEDO
abortiranje,N+c+n+Domena=MED+DomenaType=PROC+FLX=CRTANJE
abrazija,N+c+f+FLX=MEDO
adaptacija,N+c+f+FLX=MEDO
administracija,N+c+f+FLX=MEDO
admiral,N+c+m+FLX=ASTRONOM
adolescencija,N+c+f+FLX=MEDO
adolescent,N+c+m+FLX=AGENT
adoptiranje,N+c+n+FLX=CRTANJE
adoracija,N+c+f+FLX=MEDO
adoriranje,N+c+n+FLX=CRTANJE
```
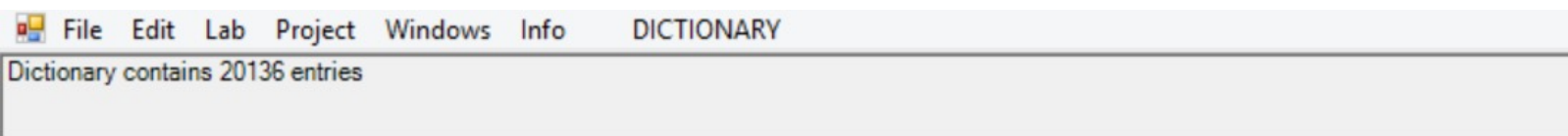
Croatian DIctionary (Kristina Kocijan, Univ. of Zagreb)

File   Edit   Lab   Project   Windows   Info   DICTIONARY

Dictionary contains 20136 entries

```
#use hrCroNoun.nof

#IMENICE

fakultet,N+c+m+FLX=ALAT

abdiciranje,N
abdikacija,N+
abdukcija,N+c
abeceda,N+c+f
aberacija,N+c
abolicija,N+c
abortiranje,N
abrazija,N+c+
adaptacija,N+
administracij
admiral,N+c+m
adolescencija
adolescent,N+
adoptiranje,N
adoracija,N+c
adoriranje,N+
```

File   Edit   La   Project   Windows   Info
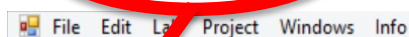
```
ALAT      = :n021s + :n021p;

MEDO         = :n50s + :n65p;

#embedded graphs:
n021s = <E>/Nom+s + a/G+s + u/D+s + <E>/Acc+s + e/Voc+s + u/L+s + om/I+s;
n021p = i/Nom+p + a/G+p + ima/D+p + e/Acc+p + i/Voc+p + ima/L+p + ima/I+p;
n50s    = <E>/Nom+s + <B1>(e/G+s + i/D+s + u/Acc+s + o/Voc+s + i/L+s + om/I+s);
n65p    = <B1>(e/Nom+p + a/G+p + ama/D+p + e/Acc+p + e/Voc+p + ama/L+p + ama/I+

n060s   = <E>/Nom+s + a/G+s + u/D+s + <E>/Acc+s + <B1>če/Voc+s + u/L+s + om/I+s
n0323s = <E>/Nom+s + a/G+s + u/D+s + <E>/Acc+s + <E>/Voc+s + u/L+s + om/I+s;
n041s  = <E>/Nom+s + a/G+s + u/D+s + <E>/Acc+s + u/Voc+s + u/L+s + em/I+s;
n025s  = <E>/Nom+s + a/G+s + u/D+s + <E>/Acc+s + u/Voc+s + u/L+s + om/I+s;
n0812s = <E>/Nom+s + a/G+s + u/D+s + <E>/Acc+s + u/Voc+s + <B1>že/Voc+s + u/L+
n022s  = <E>/Nom+s + a/G+s + u/D+s + a/Acc+s + e/Voc+s + u/L+s + om/I+s;
n02s   = <E>/Nom+s + a/G+s + u/D+s + a/Acc+s + <E>/Voc+s + u/L+s + om/I+s;
n001s  = <E>/Nom+s + a/G+s + u/D+s + a/Acc+s + <B1>če/Voc+s + u/L+s + om/I+s;
```

NooJ Dictionaries
Phrasal Verbs, Peter Machonis, International Florida University

NooJ - [C:\Users\max\Documents\NooJ\en\Lexical Analysis_phrasal verbs.dic]

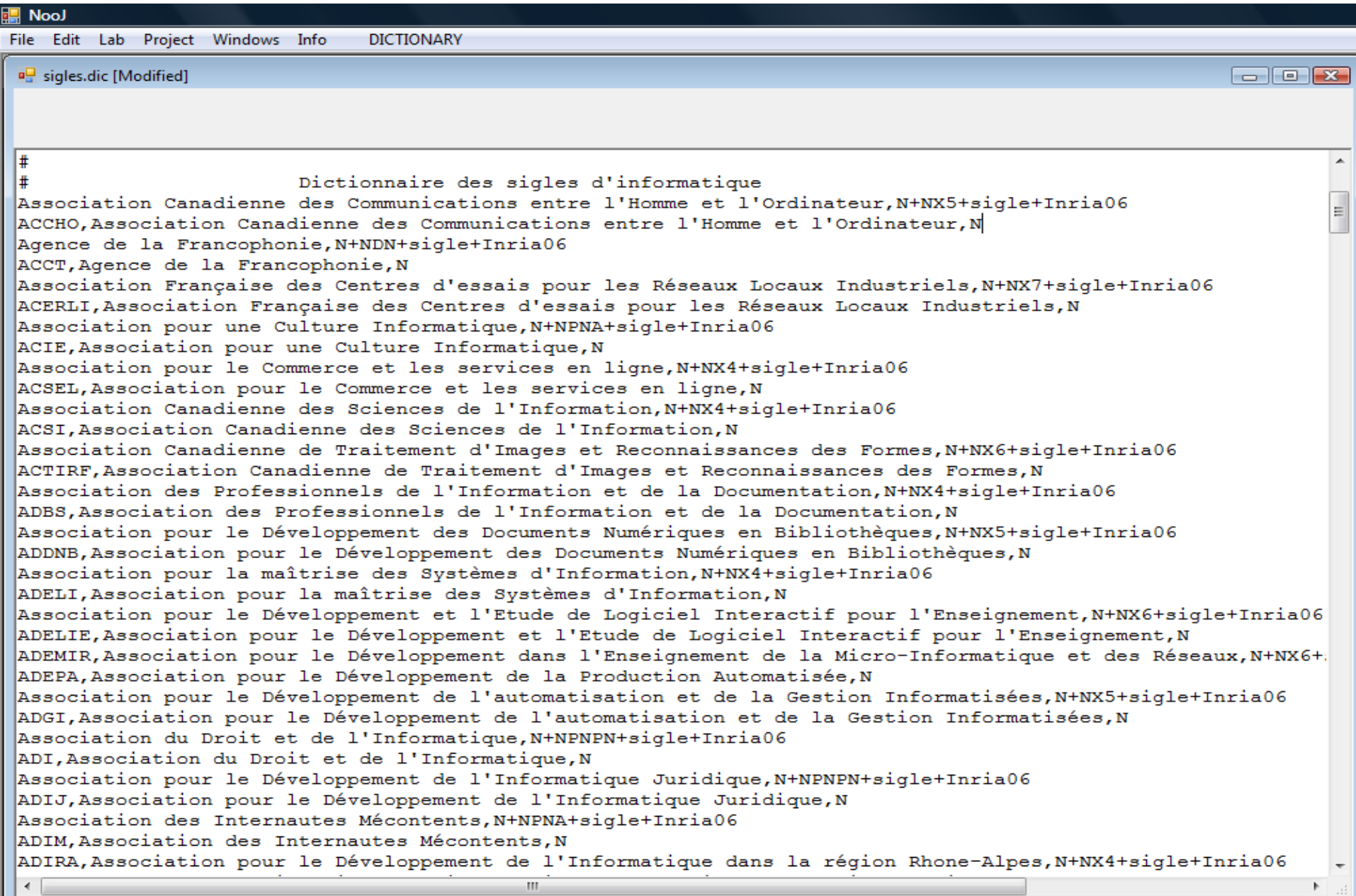File   Edit   Lab   Project   Windows   Info   DICTIONARY

Dictionary contains 1260 entries

Entry	Part	FLX	Nom	N1example	synonym	FR
blank	out	ASK		the paragraph	cover over	
blast	out	ASK		the music	blast	
block	up	ASK	block up	the sink	obstruct	obstruer
block	off	ASK		the street	close	
block	out	ASK		the play	sketch	
block	out	ASK		sensitive mat...	obscure	
block	out	ASK		the memories	suppress	
blot	out	ADMIT		the answer	erase	
blot	out	ADMIT		the bad memories	suppress	
blow	up	BLOW	blowup of N1	the photo	enlarge	agrandir
blow	up	BLOW	blow up	the scandal	exaggerate	exagérer
blow	up	BLOW		the building	explode	exploser
blow	up	BLOW		the balloons	inflate	gonfler
blow	away	BLOW		the spectators	overwhelm	
blow	away	BLOW		the leaves	blow	
blow	down	BLOW		the tree	cause to fall	
blow	off	BLOW		class	not attend or accompany	
blow	out	BLOW		the candle	extinguish	
blow	out	BLOW	have a blowout	the front tire	deflate suddenly	
blow	out	BLOW		the windows	break	
blow	out	BLOW	it was a blow-out	the other team	easily defeat	
blow	over	BLOW		the tree	cause to fall	
bluff	out	ASK		the concierge	deceive or lie about something	
blurt	out	ASK		the information	say something without thinking	
bog	down	ADMIT		the peace talks	delay progress of	
boil	up	ASK		some water	boil	bouillir
boil	away	ASK		the water	boil	
boil	down	ASK		the mushrooms	boil	
bolster	up	ASK		Max	give hope to	donner espoir
bolster	up	ASK		the theory	support	soutenir
boom	out	ASK		the music	play loudly	
boot	up	ASK		the computer	start	démarrer

Cancel

NooJ Dictionaries
Organizations, Farida Aoughlis, Mouloud Mammeri University (Algeria)


```
#
#                    Dictionnaire des sigles d'informatique
Association Canadienne des Communications entre l'Homme et l'Ordinateur,N+NX5+sigle+Inria06
ACCHO,Association Canadienne des Communications entre l'Homme et l'Ordinateur,N
Agence de la Francophonie,N+NDN+sigle+Inria06
ACCT,Agence de la Francophonie,N
Association Française des Centres d'essais pour les Réseaux Locaux Industriels,N+NX7+sigle+Inria06
ACERLI,Association Française des Centres d'essais pour les Réseaux Locaux Industriels,N
Association pour une Culture Informatique,N+NPNA+sigle+Inria06
ACIE,Association pour une Culture Informatique,N
Association pour le Commerce et les services en ligne,N+NX4+sigle+Inria06
ACSEL,Association pour le Commerce et les services en ligne,N
Association Canadienne des Sciences de l'Information,N+NX4+sigle+Inria06
ACSI,Association Canadienne des Sciences de l'Information,N
Association Canadienne de Traitement d'Images et Reconnaissances des Formes,N+NX6+sigle+Inria06
ACTIRF,Association Canadienne de Traitement d'Images et Reconnaissances des Formes,N
Association des Professionnels de l'Information et de la Documentation,N+NX4+sigle+Inria06
ADBS,Association des Professionnels de l'Information et de la Documentation,N
Association pour le Développement des Documents Numériques en Bibliothèques,N+NX5+sigle+Inria06
ADDNB,Association pour le Développement des Documents Numériques en Bibliothèques,N
Association pour la maîtrise des Systèmes d'Information,N+NX4+sigle+Inria06
ADELI,Association pour la maîtrise des Systèmes d'Information,N
Association pour le Développement et l'Etude de Logiciel Interactif pour l'Enseignement,N+NX6+sigle+Inria06
ADELIE,Association pour le Développement et l'Etude de Logiciel Interactif pour l'Enseignement,N
ADEMIR,Association pour le Développement dans l'Enseignement de la Micro-Informatique et des Réseaux,N+NX6+.
ADEPA,Association pour le Développement de la Production Automatisée,N
Association pour le Développement de l'automatisation et de la Gestion Informatisées,N+NX5+sigle+Inria06
ADGI,Association pour le Développement de l'automatisation et de la Gestion Informatisées,N
Association du Droit et de l'Informatique,N+NPNPN+sigle+Inria06
ADI,Association du Droit et de l'Informatique,N
Association pour le Développement de l'Informatique Juridique,N+NPNPN+sigle+Inria06
ADIJ,Association pour le Développement de l'Informatique Juridique,N
Association des Internautes Mécontents,N+NPNA+sigle+Inria06
ADIM,Association des Internautes Mécontents,N
ADIRA,Association pour le Développement de l'Informatique dans la région Rhone-Alpes,N+NX4+sigle+Inria06
```

Exercise

- Construct the dictionary that covers the vocabylary of the text:

  That boy thinks that she's out there.

Exercise

- Construct the dictionary that covers the vocabylary of the text:

  That boy thinks that she's out there.

- *that* (ADV, CONJ or DET), *boy* (INTJ or NOUN), *think* (VERB), *she* (PRO), *'s* (GEN), *be* (VERB), *out* (PART or PREP) *there* (ADV), *out there* (ADV+UNAMB)

Inflectional Morphology

- A set of rules = a grammar
- Each rule describes an inflectional paradigm, *i.e.,* how to compute the set of inflected forms and their properties for a given set of lexical entries:

```
Paradigm = Suffix/Information |
Suffix/Information | Suffix/Information … ;
```

Lab > Morphology

Inflectional Morphology

- Each rule describes an inflectional paradigm:

  **Paradigm = Suffix/Information |
  Suffix/Information | Suffix/Information** … **;**

- Suffixes can contain operators:

  <E>: Empty suffix, **<B>**: Backspace, **<D>**: Duplicate,

  <L>: Left, **<R>**: Right,

  <Á>: Add Acute Accent, **<A>**: Remove Accent,

  <N>: Next wordform, **<P>**: Previous wordform,

  <F>: Finalize last consonant…

Lab > Morphology

Exercise

- Construct the inflectional grammar rule for the verb *to help*

Exercise

- Construct the inflectional grammar rule for the verb *to help*

HELP = <E>/STANDARD | s/PRESENT+3rd+sing | ing/GERUND | ed/PASTPARTICIPLE | ed/PRETERIT ;
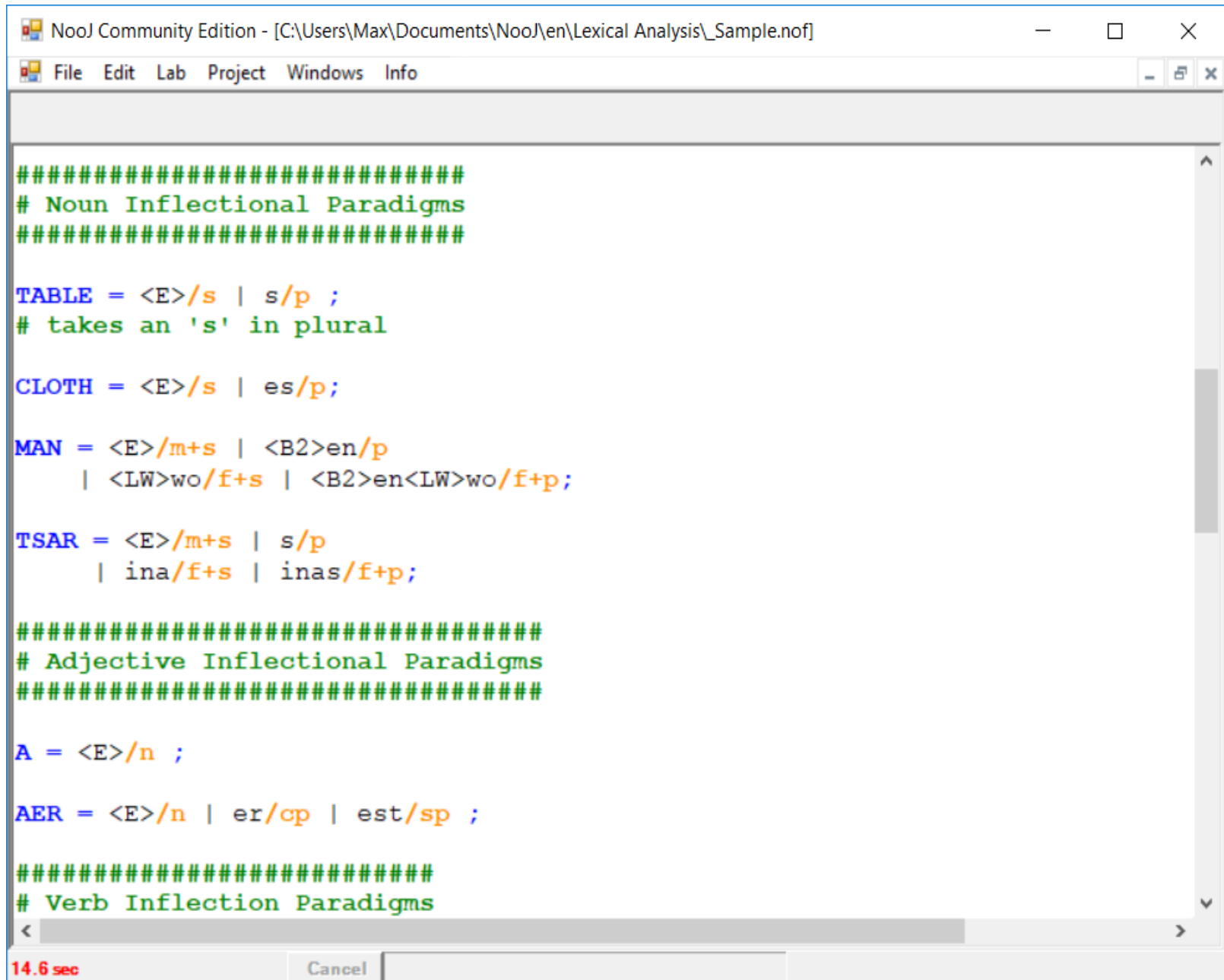
Exercise

- Construct the inflectional grammar rule for the verb *to love*

Exercise

- Construct the inflectional grammar rule for the verb *to love*

LOVE = <E>/STANDARD | s/PRESENT+3rd+sing | <B>ing/GERUND | ed/PASTPARTICIPLE | ed/PRETERIT ;

Inflectional Morphology

```
NooJ Community Edition - [C:\Users\Max\Documents\NooJ\en\Lexical Analysis\_Sample.nof]    —    □    ✕

 File   Edit   Lab   Project   Windows   Info                                              _  ⬚  ✕


##############################
# Noun Inflectional Paradigms
##############################

TABLE = <E>/s | s/p ;
# takes an 's' in plural

CLOTH = <E>/s | es/p;

MAN = <E>/m+s | <B2>en/p
    | <LW>wo/f+s | <B2>en<LW>wo/f+p;

TSAR = <E>/m+s | s/p
     | ina/f+s | inas/f+p;

##################################
# Adjective Inflectional Paradigms
##################################

A = <E>/n ;

AER = <E>/n | er/cp | est/sp ;

############################
# Verb Inflection Paradigms
```

14.6 sec   Cancel

Inflectional Morphology


```
NooJ Community Edition - [C:\Users\Max\Documents\NooJ\en\Lexical Analysis\_Sample.nof]        —    □    ×

File   Edit   Lab   Project   Windows   Info                                                    _ ⊡ ×


##############################
# Verb Inflection Paradigms
##############################


BE = <E>/INF | ing/G | en/PP
   | <BW> (am/PR+s+1 | are/PR+s+2 | is/PR+s+3 | are/PR+p+1 | are/PR+p+2 | ar
   | <BW> (was/PT+s+1 | were/PT+s+2 | was/PT+s+3 | were/PT+p+1 | were/PT+p+2


DO = <E>/INF | ing/G | ne/PP
   | <E>/PR+s+1 | <E>/PR+s+2 | es/PR+s+3 | <E>/PR+p+1 | <E>/PR+p+2 | <E>/PR+
   | <B> (id/PT+s+1 | id/PT+s+2 | id/PT+s+3 | id/PR+p+1 | id/PR+p+2 | id/PR+


ASK = <E>/INF | ed/PP | ing/G
   | <E>/PR+1+2+s | <E>/PR+1+2+3+p | s/PR+3+s
   | ed/PT+1+2+3+s+p;


HAVE = <E>/INF | <B>ing/G | <B2>d/PP
   | <E>/PR+s+1 | <E>/PR+s+2 | <B2>s/PR+s+3 | <E>/PR+p+1 | <E>/PR+p+2 | <E
   | <B2> (d/PT+s+1 | d/PT+s+2 | d/PT+s+3 | d/PR+p+1 | d/PR+p+2 | d/PR+p+3


EAT = <E>/INF | ing/G | en/PP
   | <E>/PR+s+1 | <E>/PR+s+2 | s/PR+s+3 | <E>/PR+p+1 | <E>/PR+p+2 | <E>/PR+
   | <BW> (ate/PT+s+1 | ate/PT+s+2 | ate/PT+s+3 | ate/PR+p+1 | ate/PR+p+2 |

HELP = <E>/INF | ing/G | ed/PP

14.6 sec              Cancel
```
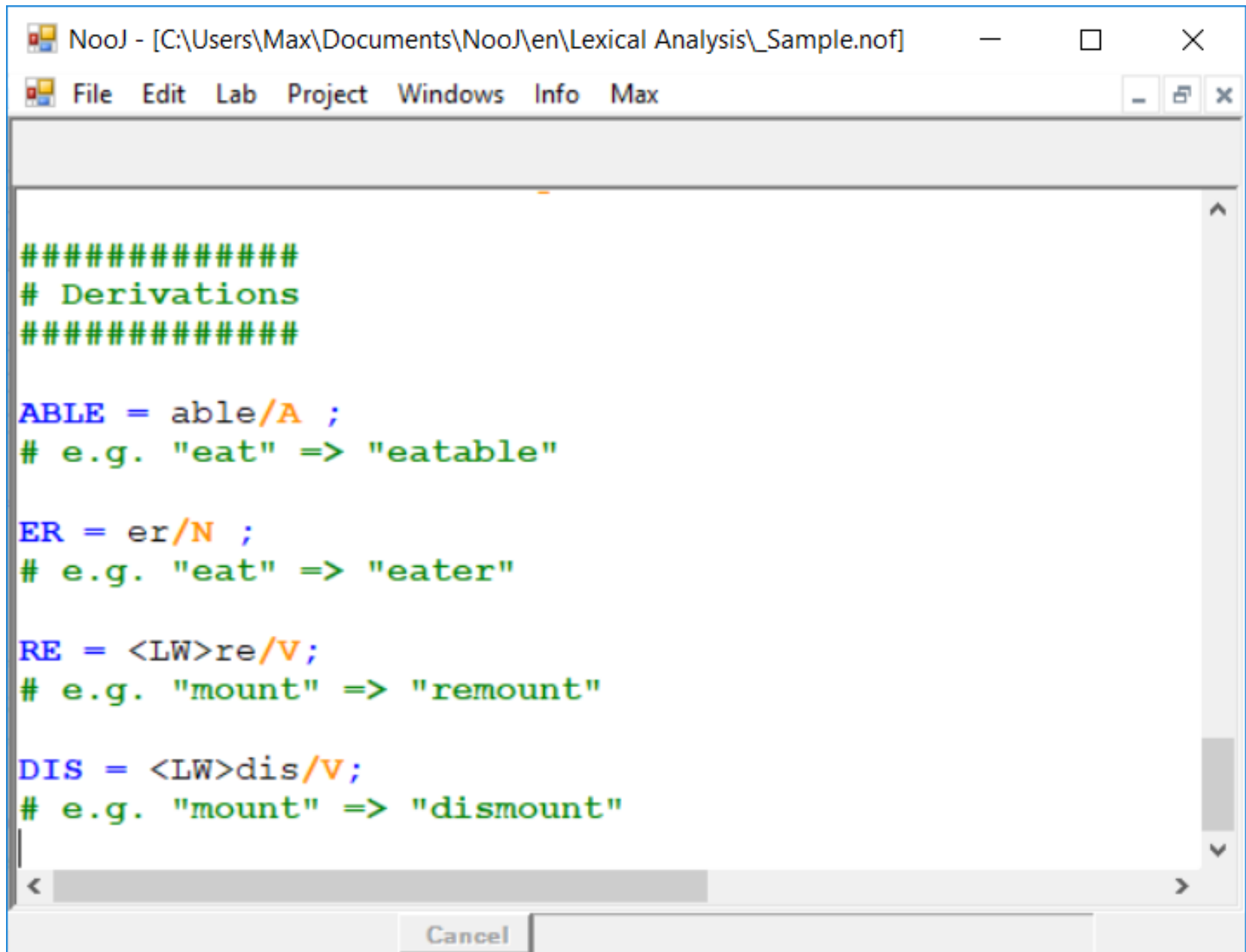
Derivational Morphology

File   Edit   Lab   Project   Windows   Info   Max

```
#############
# Derivations
#############


ABLE = able/A ;
# e.g. "eat" => "eatable"


ER = er/N ;
# e.g. "eat" => "eater"


RE = <LW>re/V;
# e.g. "mount" => "remount"


DIS = <LW>dis/V;
# e.g. "mount" => "dismount"
```

Cancel

Inflectional and Derivational Description

do,V+FLX=DO+DRV=ABLE:<E>+DRV=ER:TABLE+DRV=RE

FLX=DO: produces the forms *do, does, doing, done*

DRV=ABLE:<E>: produce the adjective *doable*

DRV=ER:TABLE: produce the nominal forms *doer* and *doers*

DRV=RE: produce the verbal forms *redo, redoes, redoing, redone*

All wordforms: *do, does, doing, done, doable, doer, doers, redo, redoes, redoing, redone* are linked to the ALU *do*

CONGRATULATIONS

You know how to construct your own dictionary, and associate each of its entries with its associated inflected and derived forms