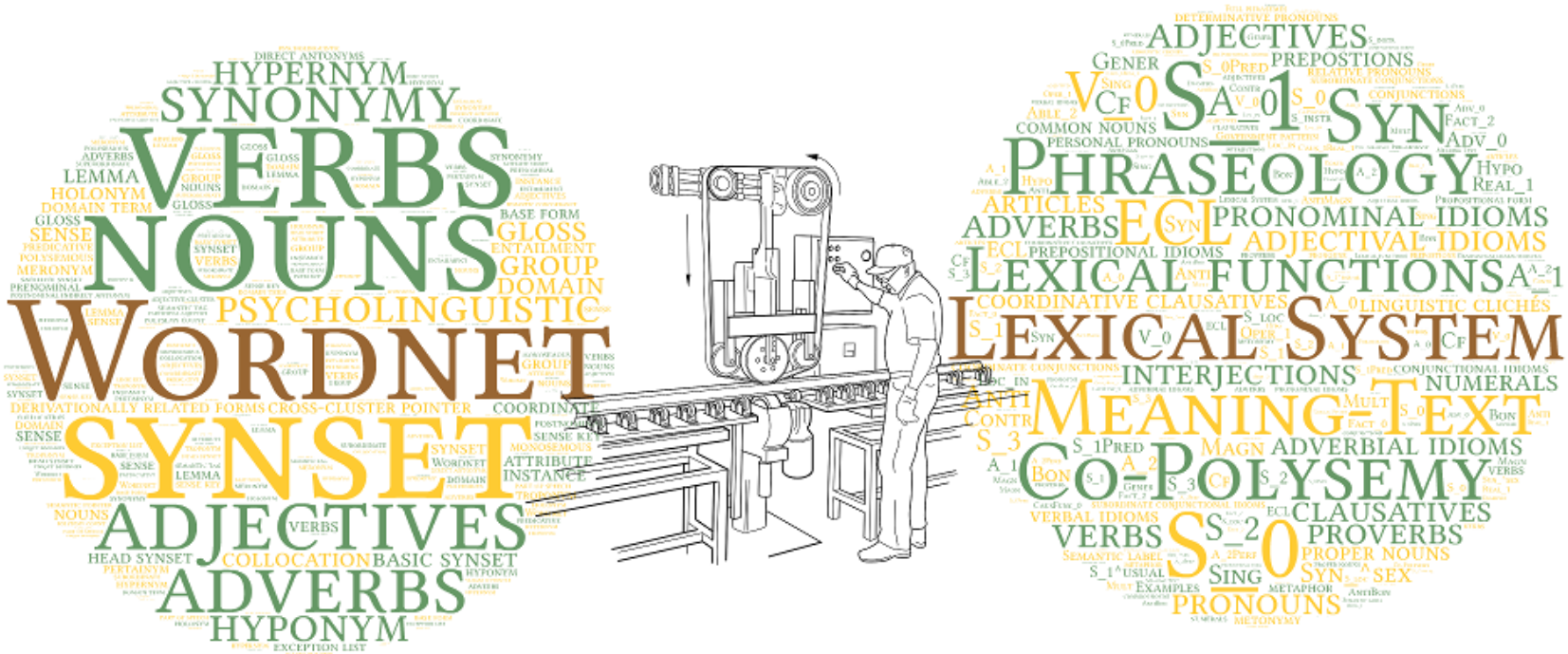
One Lexicon, Two Structures: So What Gives?

Nabil Gader, Sandrine Ollinger, Alain Polguère

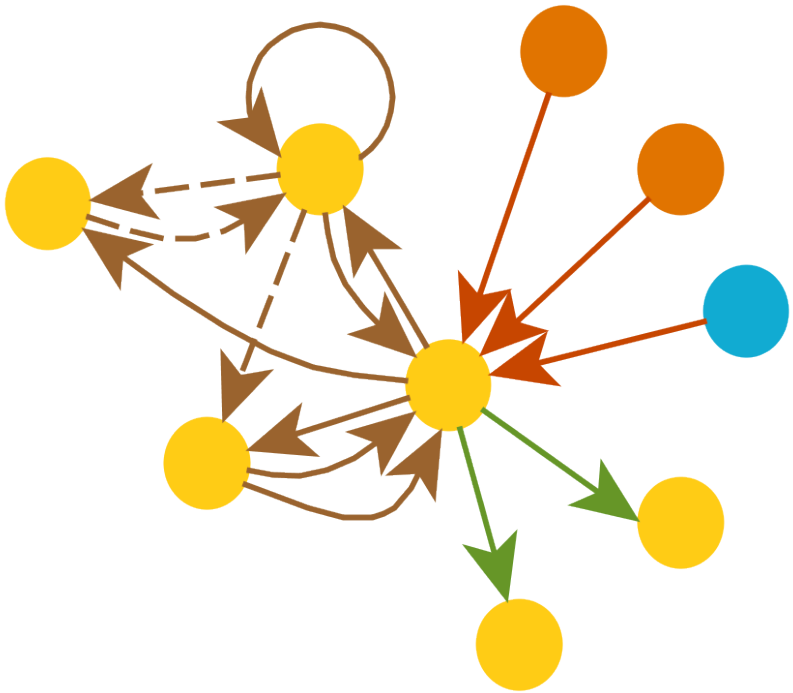
ATILF - Nancy

29 January 2014

Transmute WordNet into Lexical system



Lexical System structure



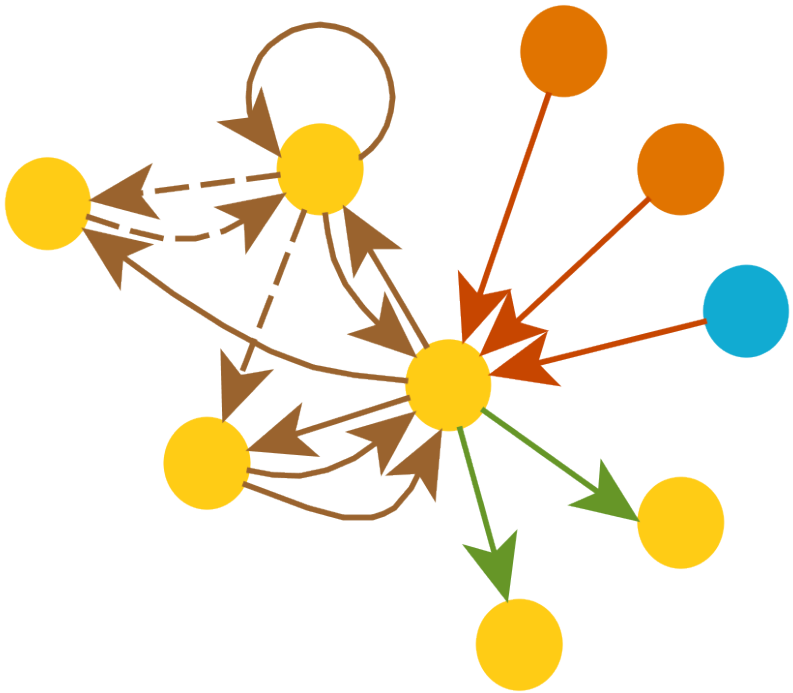
Nodes

- Lexemes
- Idioms
- Linguistic cliches

Arcs

- Lexical functions links
- Phraseology links
- Copolysemy links

Lexical System structure



Nodes

- Lexemes
- Idioms
- Linguistic cliches

Arcs

- Lexical functions links
- Phraseology links
- Copolysemy links

Lexicographic article associated with each lexical unit :
GC, DF, GP, LF, EX, PH

Confidence index

Lexical system by example : fr-LN

Nodes

- 21,507 nodes
  - 55% N
  - 21% V
  - 16% Adj
  - 0,5% Adv
- 16% single nodes

Arcs

- 39,777 arcs
  - 88% LF links
    - 1/5 syntagmatic
    - 4/5 paradigmatic
  - 10% phraseology links
  - 2% copolysemy links
- 39 loops
- 515 multiples
- 18,530 mutuals

Lexical system by example : fr-LN

Nodes

- 21,507 nodes
  - 55% N
  - 21% V
  - 16% Adj
  - 0,5% Adv
- 16% single nodes

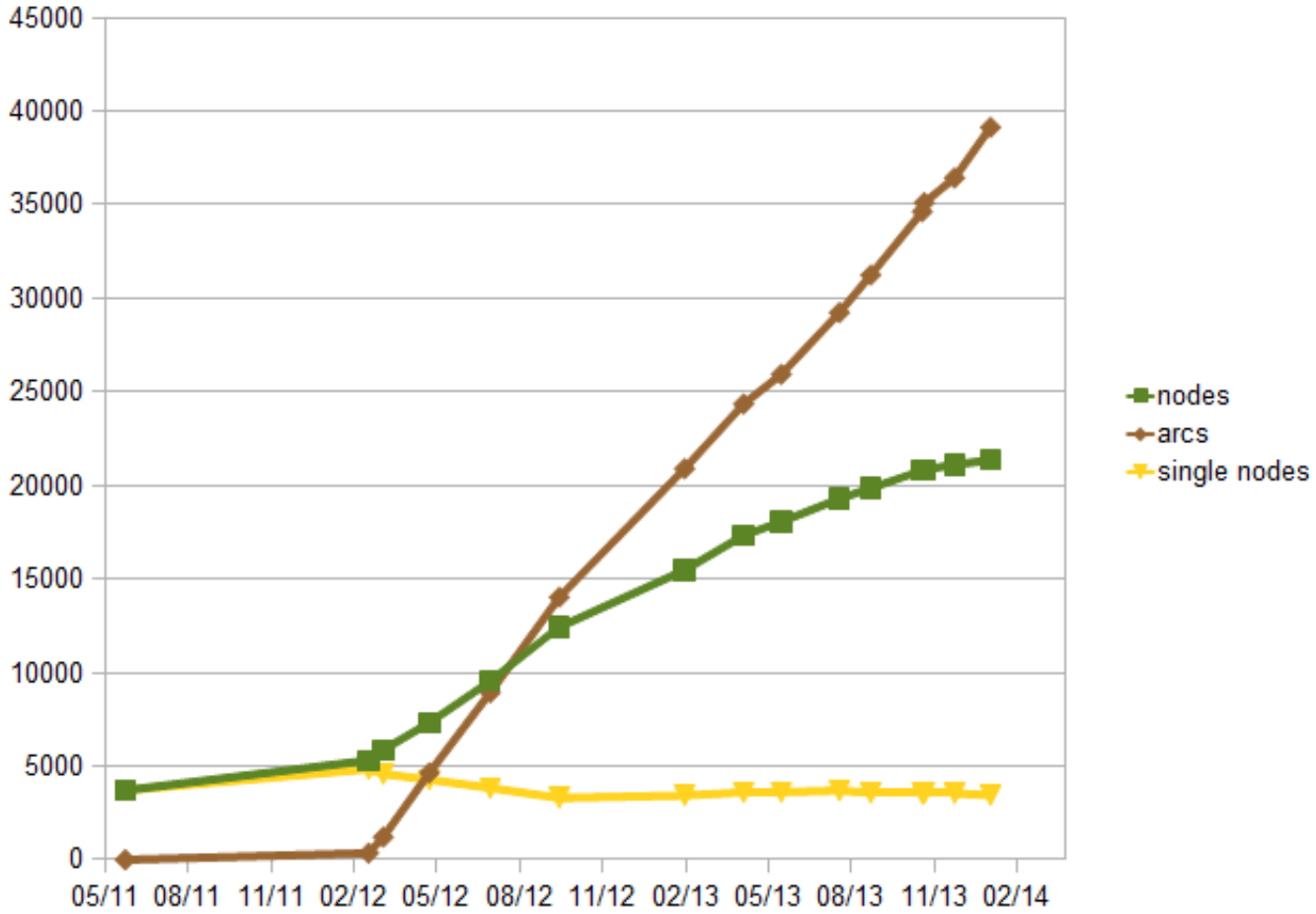
Arcs

- 39,777 arcs
  - 88% LF links
    - 1/5 syntagmatic
    - 4/5 paradigmatic
  - 10% phraseology links
  - 2% copolysemy links
- 39 loops
- 515 multiples
- 18,530 mutuals

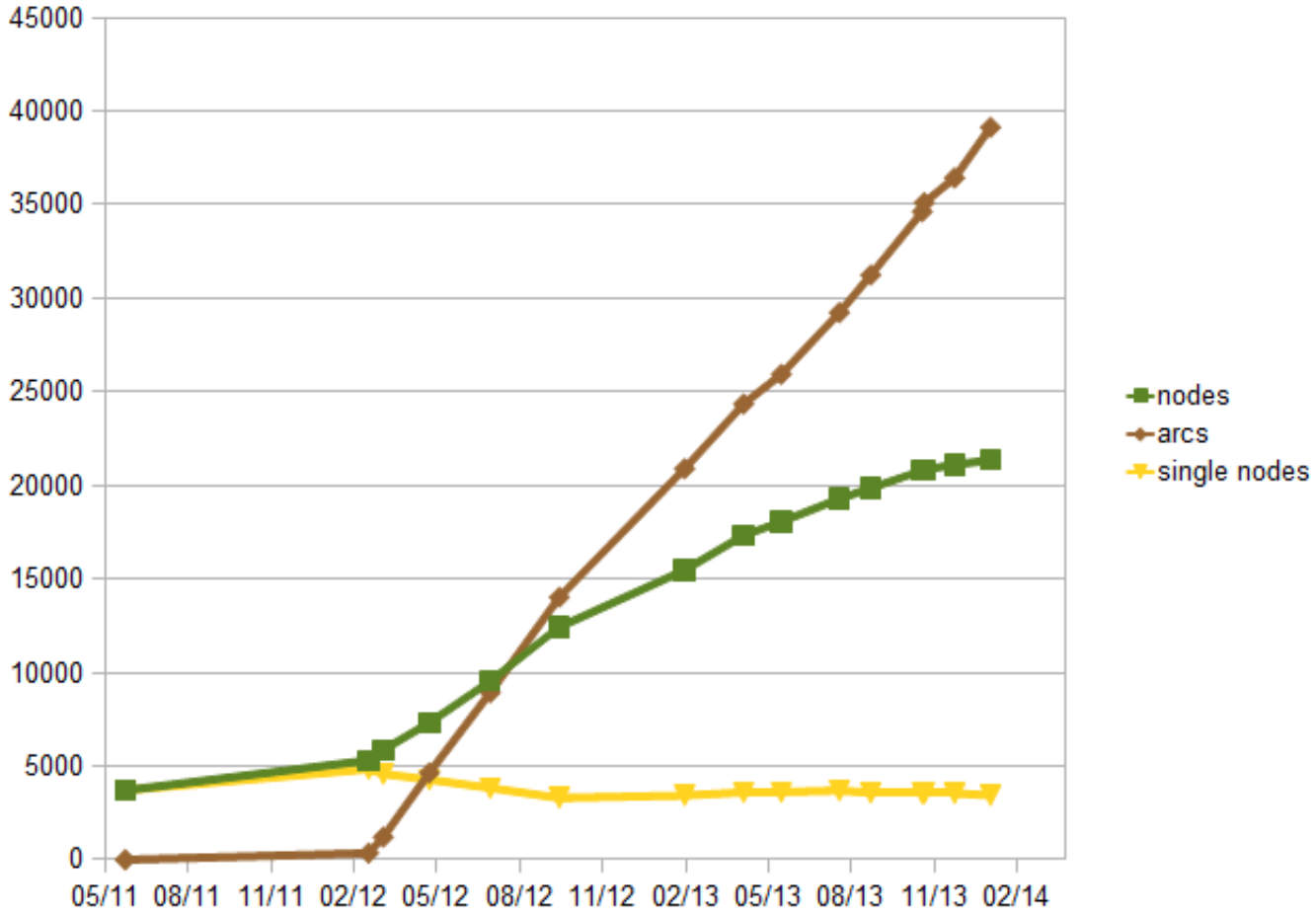
Workforce

10 full-time lexicographers since June 2011

Evolution of fr-LN



Evolution of fr-LN



Tends to become Hierarchical Small World Network

Why "transmute" WordNet into LN system ?

Start work on the English Lexicon

Explore structural behavior of LN

How to "transmute" WordNet 3.0 into en-LN ?

Nomenclature

How to "transmute" WordNet 3.0 into en-LN ?

Nomenclature

wn_s.pl
wn_g.pl
wn_sk.pl
wn_syntax.pl

How to "transmute" WordNet 3.0 into en-LN ?

Nomenclature

wn_s.pl
wn_g.pl
wn_sk.pl
wn_syntax.pl

156,584 en-LN vocables

idVoc	form	subscript

subscript : *null*,N,V,Adj,Adv

How to "transmute" WordNet 3.0 into en-LN ?

Nomenclature

wn_s.pl
wn_g.pl
wn_sk.pl
wn_syntax.pl

156,584 en-LN vocables

idVoc	form	subscript

subscript : *null*,N,V,Adj,Adv

206,976 en-LN senses

idSense	idVoc	sense number	gloss

How to "transmute" WordNet into en-LN ?

Relations

wn_s.pl
wn_hyp.pl
wn_ins.pl
wn_sim.pl
wn_mm.pl
wn_ms.pl
wn_mp.pl
wn_der.pl
wn_at.pl
wn_per.pl
wn_cs.pl
wn_ant.pl
wn_ppl.pl

How to "transmute" WordNet into en-LN ?

Relations

wn_s.pl
wn_hyp.pl
wn_ins.pl
wn_sim.pl
wn_mm.pl
wn_ms.pl
wn_mp.pl
wn_der.pl
wn_at.pl
wn_per.pl
wn_cs.pl
wn_ant.pl
wn_ppl.pl

946,208 en-LN LF links

idLF	idSenseSource	idSenseTarget

11 pre-existent LF
1 new LF : **Unspecified derivative**

How to "transmute" WordNet into en-LN ?

Lexicographic article

wn_s.pl
wn_fr.pl

How to "transmute" WordNet into en-LN ?

Lexicographic article

wn_s.pl
wn_fr.pl

en-LN senses GC

idSense	idGC	form?

How to "transmute" WordNet into en-LN ?

Lexicographic article

wn_s.pl
wn_fr.pl

en-LN senses GC

idSense	idGC	form?

en-LN senses GP comments

idSense	subcategorization frames

Future

Blue-print of an English lexical network

- Adjust actual links
- Increase connectivity
  - LF
  - Co-polysemy
  - Phraselogy
- Complete lexicographic descriptions

Tools

En-LN is compatible with our lexicographic editor


Thanks for your attention

Pedigree of th fr-/en-LNs (global)

	fr-LN	en-LN
n	21,507	206,976
m	39,777	946,208
<k>	3.6406	5.9029
Directed	true	true
Mutuals	18,530	942,795
Loops	39	1
Single	3,404	19,756
Multiples	515	124
ncc	14,013	34,342
C	0.1270	0.1031
Out degree distribution		
a	-2,0454	-1.8479
r^2	0.9541	0.8453

	C Random
fr-LN	0.00017
en-LN	0.00004

Pedigree of th fr-/en-LNs (LCC)

	fr-LN	en-LN
n_lcc	2,741	144,294
m_lcc	9,979	851,748
C_lcc	0.3225	0.0980
L_lcc	12.0383	10.1479

n_lcc	$\log n / \log\log n$
fr-LN	6.4105
en-LN	7.240

Syn∩

Wordnet Synset

s(100478262,1,'soccer',n,1,0).
s(100478262,2,'association football',n,1,0).

en-LN LF

Syn∩(SOCCER) = ASSOCIATION FOOTBALL
Syn∩(ASSOCIATION FOOTBALL) = SOCCER

Cf

Wordnet Synset

sim(301092142,301092572).

s(301092142,1,'malfunctioning',a,1,0).
s(301092142,2,'nonfunctional',a,2,0).
s(301092572,1,'bad',s,14,0).

en-LN LF

Cf(MALFUNCTIONING) = BAD_{Adj} 14
Cf(NONFUNCTIONAL) = BAD_{Adj} 14
Cf(BAD_{Adj} 14) = MALFUNCTIONING,NONFUNCTIONAL

Hypo - Gener

Wordnet Synset

hyp(110129825,110787470).

s(110129825,1,'girl',n,1,80).
s(110787470,1,'woman',n,1,143).
s(110787470,2,'adult female',n,1,0).

s(110787470) more than 15 times in wn_hyp.pl

en-LN LF

Gener(GIRL 1) = WOMAN 1,ADULT FEMALE
Hypo(WOMAN 1) = GIRL 1
Hypo(ADULT FEMALE) = GIRL 1

Hypo - Gener

Wordnet Synset

ins(109012735,108524735).

s(109012735,1,'Tartu',n,1,0).
s(108524735,1,'city',n,1,103).
s(108524735,2,'metropolis',n,1,7).
s(108524735,3,'urban center',n,1,2).

en-LN LF

Gener(TARTU) = CITY 1,METROPOLIS 1,URBAN CENTER
Hypo(CITY 1) = TARTU
Hypo(METROPOLIS 1) = TARTU
Hypo(URBAN CENTER) = TARTU

Sing - Mult

Wordnet Synset

mm(110662162,108208560).

s(110662162,1,'stringer',n,1,0).
s(108208560,1,'team',n,1,43).
s(108208560,2,'squad',n,2,4).

en-LN LF

Sing($\text{TEAM}_N\ 1$) = $\text{STRINGER}\ 1$
Sing($\text{SQUAD}\ 2$) = $\text{STRINGER}\ 1$
Mult($\text{STRINGER}\ 1$) = $\text{TEAM}_N\ 1$,$\text{SQUAD}\ 2$

Mero

Wordnet Synset

ms(101896844,103266749).

s(101896844,1,'eiderdown',n,2,0).
s(103266749,1,'eiderdown',n,1,0).
s(103266749,2,'duvet',n,1,0).
s(103266749,3,'continental quilt',n,1,0).

en-LN LF

Mero(EIDERDOWN 1) = EIDERDOWN 2
Mero(DUVET) = EIDERDOWN 2
Mero(CONTINENTAL QUILT) = EIDERDOWN 2

Mero - Holo

Wordnet Synset

mp(101896844,101853195).

s(101896844,1,'eiderdown',n,2,0).
s(101853195,1,'eider',n,1,0).
s(101853195,2,'eider duck',n,1,0).

en-LN LF

Mero(EIDER) = EIDERDOWN 2
Mero(EIDER DUCK) = EIDERDOWN 2
Holo(EIDERDOWN 2) = EIDER,EIDER DUCK

Unspecified derivative

Wordnet Synset

at(302410393,105103072).

s(302410393,1,'thick',a,1,25).
s(105103072,1,'thickness',n,1,4).

en-LN LF

Unspecified derivative($\textsc{thick}_A dj$) = $\textsc{thickness}\ 1$
Unspecified derivative($\textsc{thickness}\ 1$) = $\textsc{thick}_A dj$

Also

+ derivational morphology pointers
+ pertainym pointers

Caus

Wordnet Synset

at(202380571,202379753).

s(202380571,1,'pension off',v,1,0).
s(202379753,1,'retire',v,1,18).

en-LN LF

Caus(RETIRE 1) = //PENSION OFF 1

Anti∩

Wordnet Synset

ant(100080743,1,100080968,1).
ant(100080968,1,100080743,1).

s(100080743,1,'call option',n,2,0).
s(100080968,1,'put option',n,2,0).
s(100080968,2,'put',n,1,0).

en-LN LF

Anti∩(CALL OPTION 2) = PUT OPTION 2
Anti∩(PUT OPTION 2) = CALL OPTION 2

A_2

Wordnet Synset

ppl(303152480,1,200991683,1).

s(303152480,1,'posted',a,1,0).
s(200991683,1,'post',v,2,3).

en-LN LF

$\mathbf{A_2}(\text{POST}_V\ 2) = \text{POSTED}$

Duplicates and Co

- 5,580 word senses
- 10 derivational morphology pointers
- 3,222 pertainym pointers
- 710 antonymous pairs

- s(301380267,1,'aerial',s).
- **wrongly coded**
- no corresponding sense in our database
- 2 derivational morphology pointers lost

Troponymy between talk and whisper

en-LN LF

Gener(WHISPER) = $\text{TALK}_V\,2$, ...
Hypo($\text{TALK}_V\,2$) = WHISPER
...

corrected en-LN LF

$\textbf{Syn}_{\subset}$(WHISPER) = $\text{TALK}_V\,2$, ...
$\textbf{Syn}_{\supset}$($\text{TALK}_V\,2$) = WHISPER
AntiMagn($\text{TALK}_V\,2$) = WHISPER