Morphological acquisition by formal analogy

Jean-François Lavallée and
Philippe Langlais

DIRO, University of Montreal

Outline

- What is a formal analogy?
- System develloped
  - RALI-ANA
  - RALI-COF
- Results and analysis

Formal Analogy

Analogy **:** Relation between 4 items noted ***[X:Y::Z:T]*** which reads « X is to Y what Z is to T »
Ex: ***[*** kitten:cat::puppy:dog ***]***

Formal analogy **:** Analogy that can be identified at the graphemic level
Ex: **cordially** is to **cordial** what **appreciatively** is to **appreciative**

Formal Analogy

Formal analogy can be defined in terms of factorization (Stroppa & Yvon 2005).

$$(Y, Z) \in \{(X,T), (T, X)\}$$

cordially	=	**cordial ly**
cordial	=	**cordial ε**
appreciatively	=	**appreciative ly**
appreciative	=	**appreciative ε**

RALI-ANA System

Basic idea : Factorization computed from valid analogies correspond to morpheme boundaries and they are more frequent than fortuitous ones.

Valid				Fortuitous			
unread	**=**	**un**	**read**	**spear**	**=**	**s**	**pear**
read	**=**	**ε**	**read**	**pear**	**=**	**ε**	**pear**
undone	**=**	**un**	**done**	**swear**	**=**	**s**	**wear**
done	**=**	**ε**	**done**	**wear**	**=**	**ε**	**wear**

RALI-ANA System

Ana-seg symply picks the most frequent factorization.

abolishing (ENG)	
abolish ing	12
ab olishing	4
abol ishing	2
a bo lishing	1
abolis hing	1
abolish in g	1

RALI-ANA System

Ana-seg symply picks the most frequent factorization.

abolishing (ENG)	
abolish ing	12
ab olishing	4
abol ishing	2
a bo lishing	1
abolis hing	1
abolish in g	1

Valid

[abolishing:abolish::listening:listen]

[abolishing:abolished::looking:looked]

RALI-ANA System

Ana-seg symply picks the most frequent factorization.

abolishing (ENG)	
abolish ing	12
ab olishing	4
abol ishing	2
a bo lishing	1
abolis hing	1
abolish in g	1

Valid

[abolishing:abolish::listening:listen]

[abolishing:abolished::looking:looked]

Fortuitous

[abolishing:polishing::about:pout]

RALI-ANA System

- Submitted results are partial.
- Require all analogies to be computed
- Could take months
- For finnish, only 1/3 of the words have been analyzed

RALI-COF System

A system that generalizes learning from analogies from a subset of words computed from a subset of the lexicon .

- Requires less analogies
- Sacrifice the conceptual simplicity of RALI-ANA

CoFactors

Pair of factors occuring at the same position in the factorization of an analogy.

cordially	**=**	**cordial**	**ly**
cordial	**=**	**cordial**	**ε**
appreciatively	**=**	**appreciative**	**ly**
appreciative	**=**	**appreciative**	**ε**

Cofactors : **cordial/appreciative ε/ly.**

C-Rules

Rewritting rule derived from a cofactor which introduce the notion of context

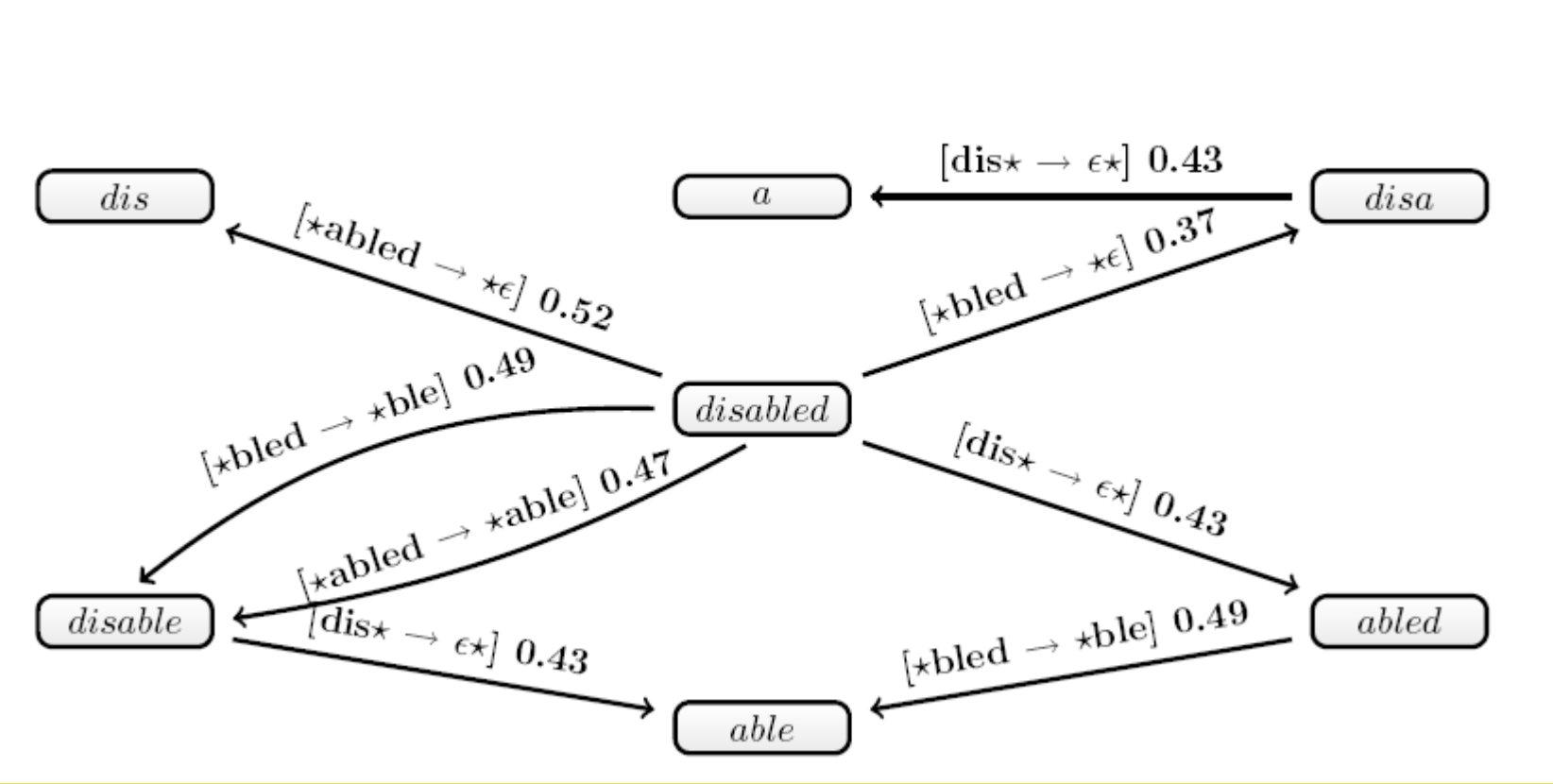
ε/ly = *ly *ε →

cordial/appreciative = appreciative* → cordial*

C-Rules are extracted from all the analogies computed.

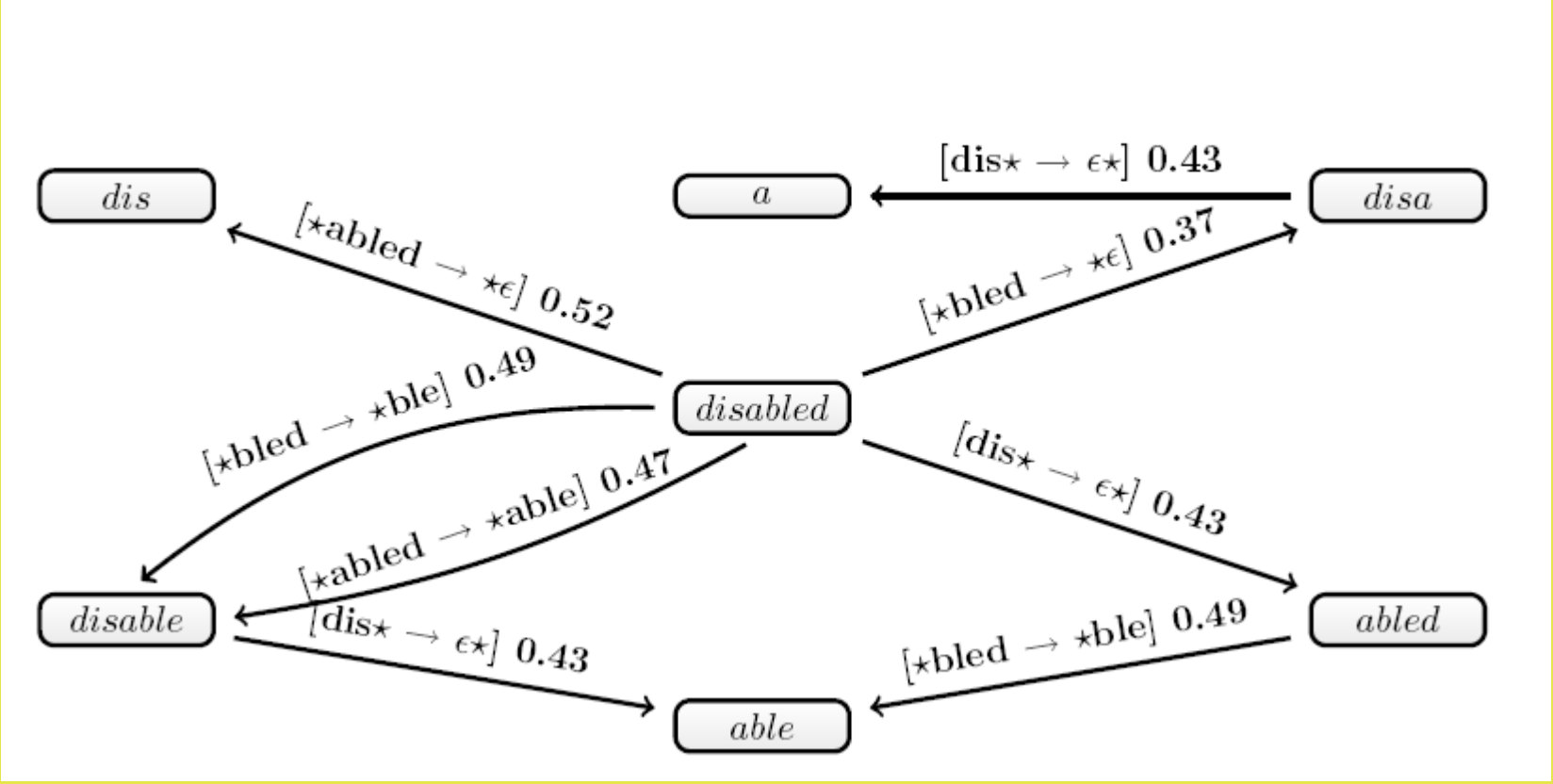
Relation Selection

A directed graph of all the forms that can be obtained from sucessive applications of C-Rule is computed for every word.



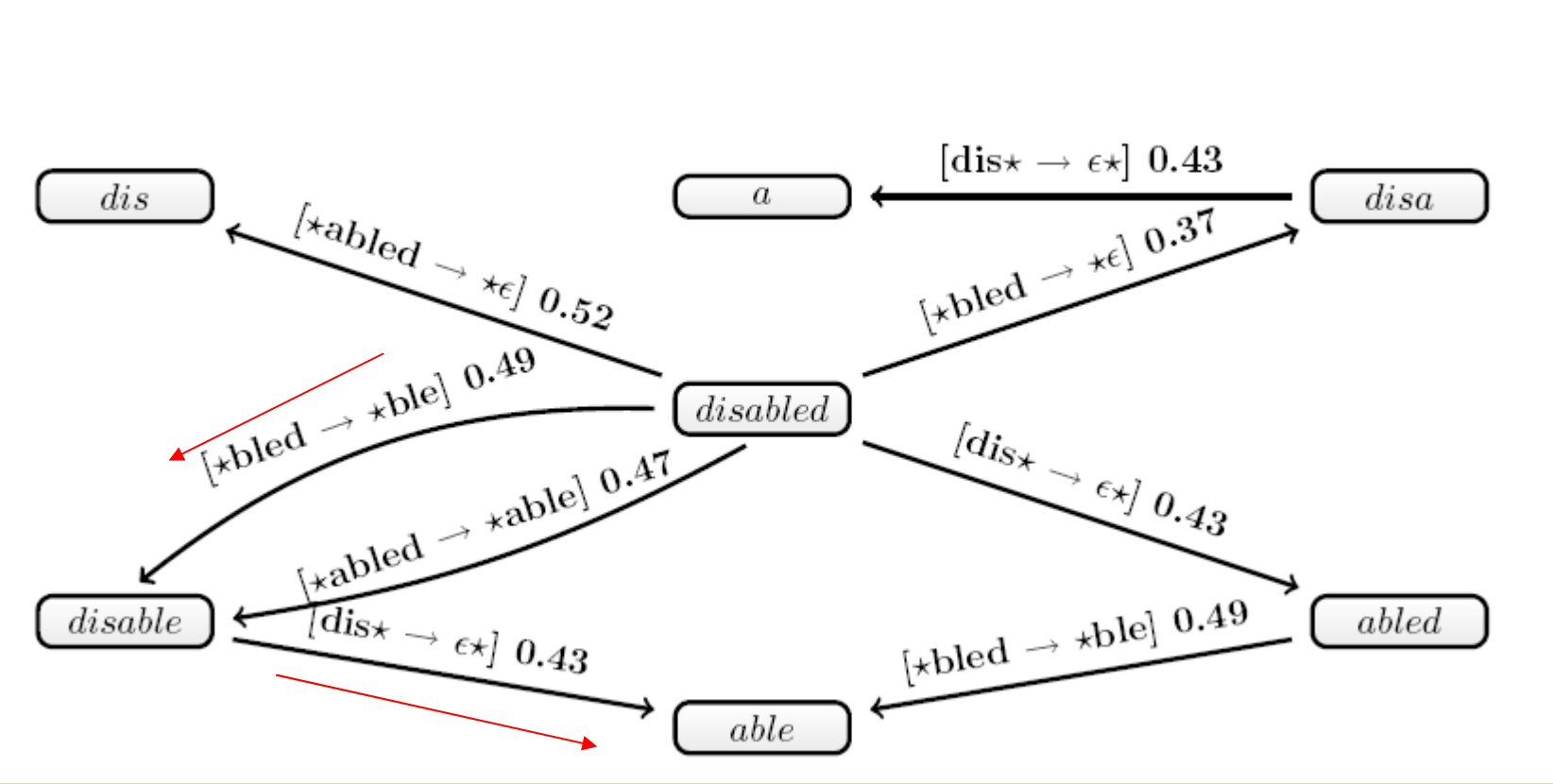
Relation Selection

A score is calculated for every form thus obtained that are present in the lexicon.



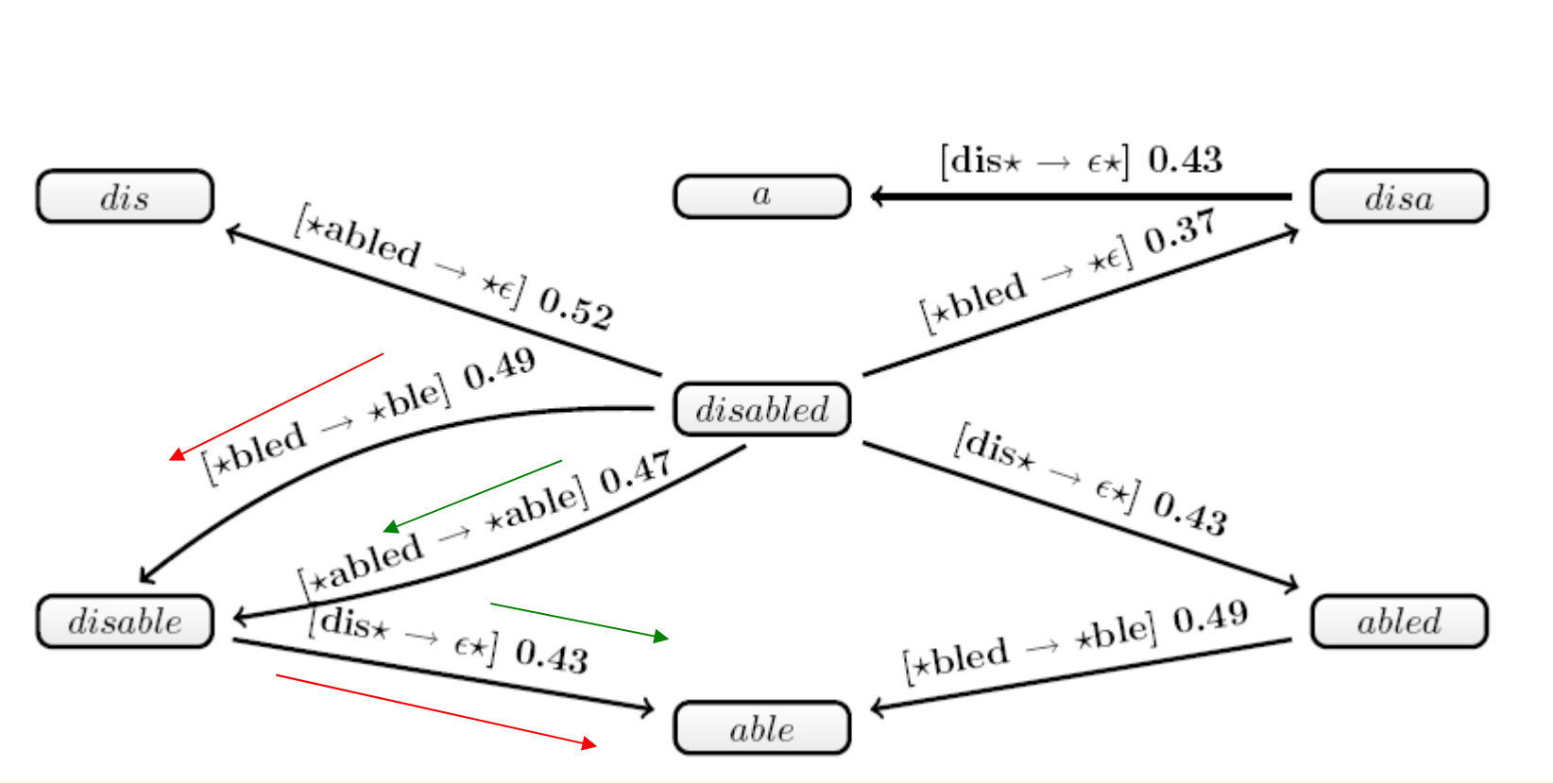
Relation Selection

disabled, able : 0.49* 0.43



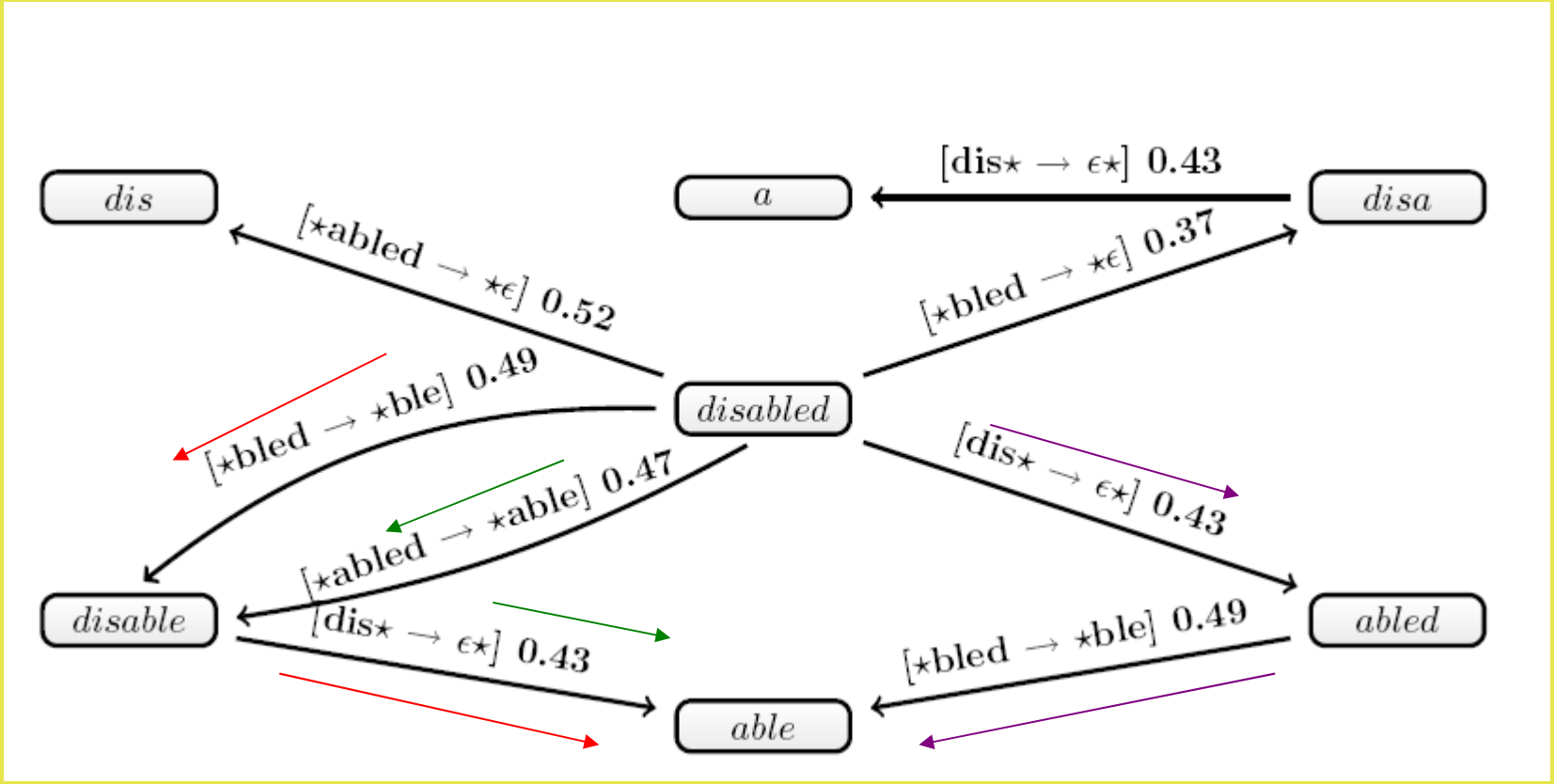
Relation Selection

disabled, able : 0.49* 0.43 + 0.47 * 0.43



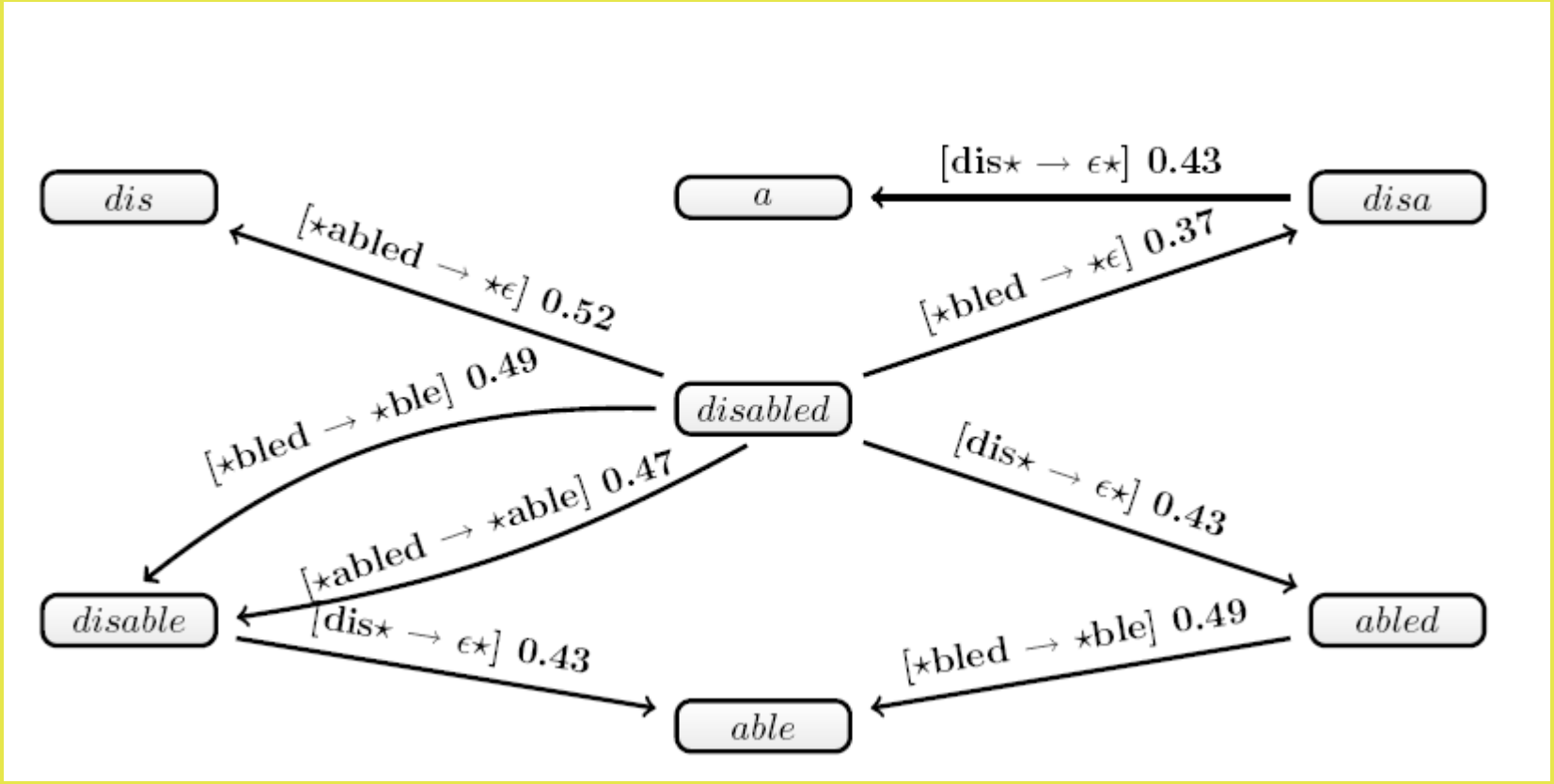
Relation Selection

disabled, able : 0.49* 0.43 + 0.47 * 0.43 + 0.43 * 0.49 = 0.62



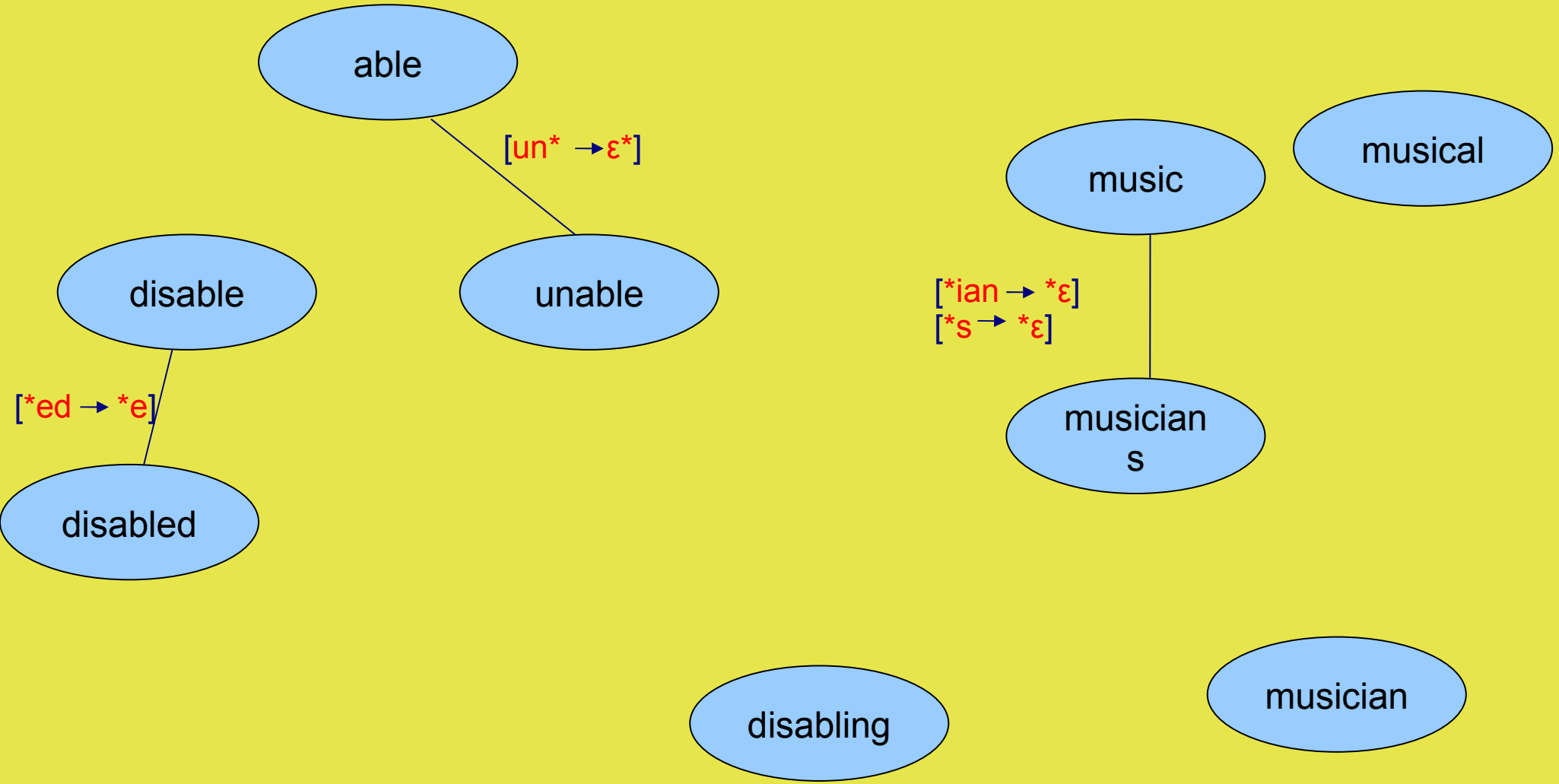
Relation Selection

disabled, dis : 0.52
disabled, able : 0.62
disabled, disable : 0.96
. . .

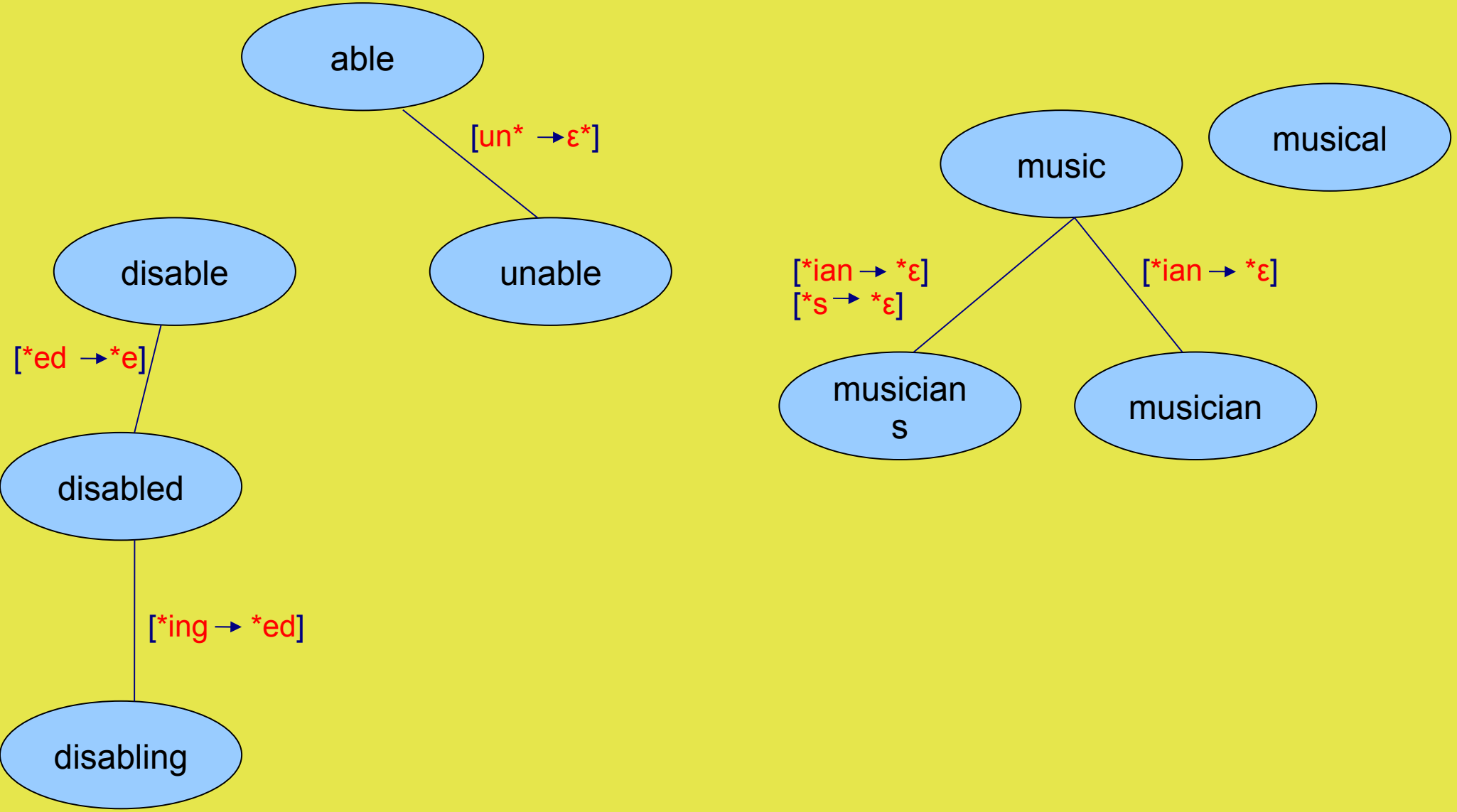


Word Relation Tree

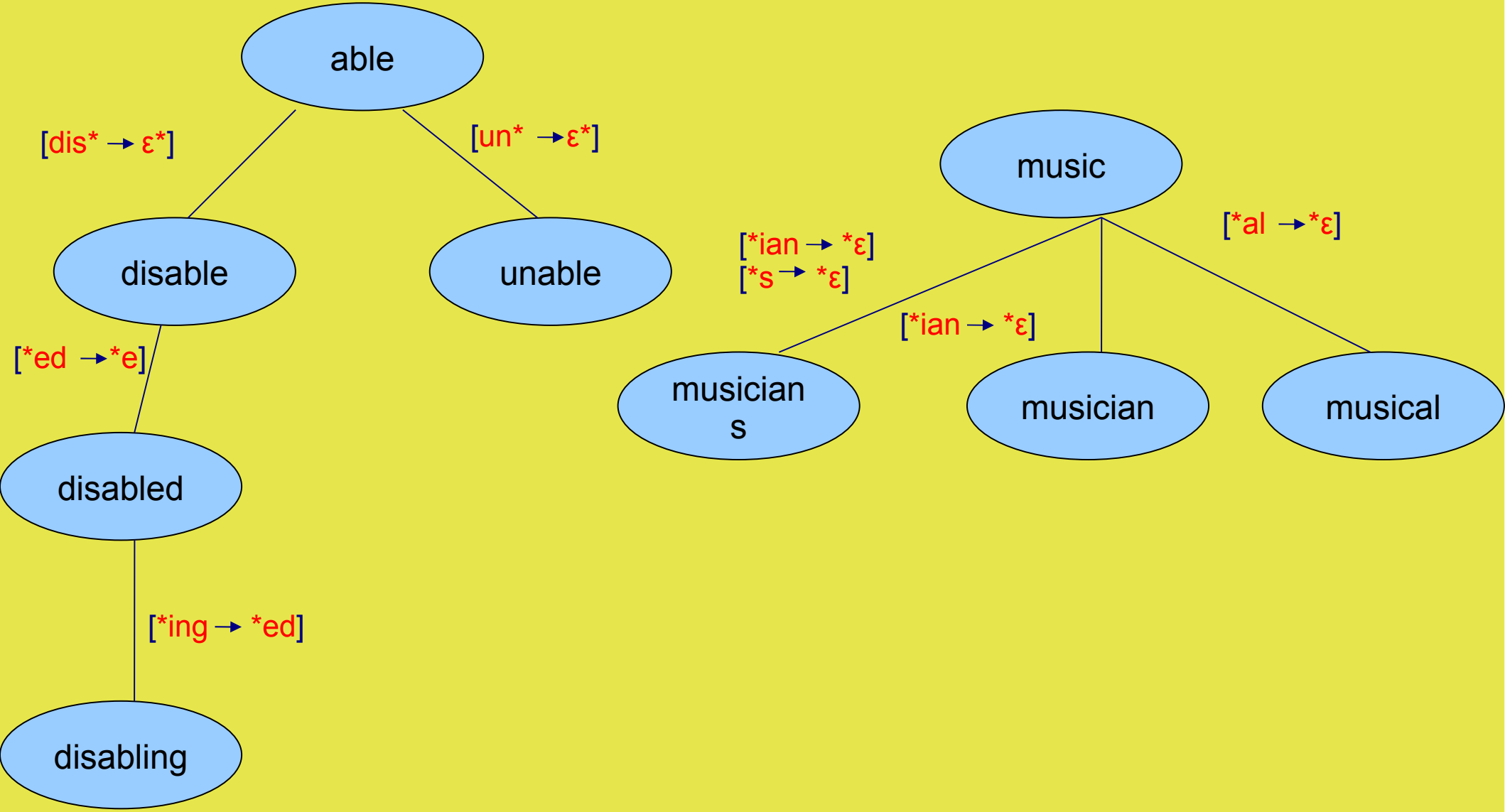
Word Relation Tree



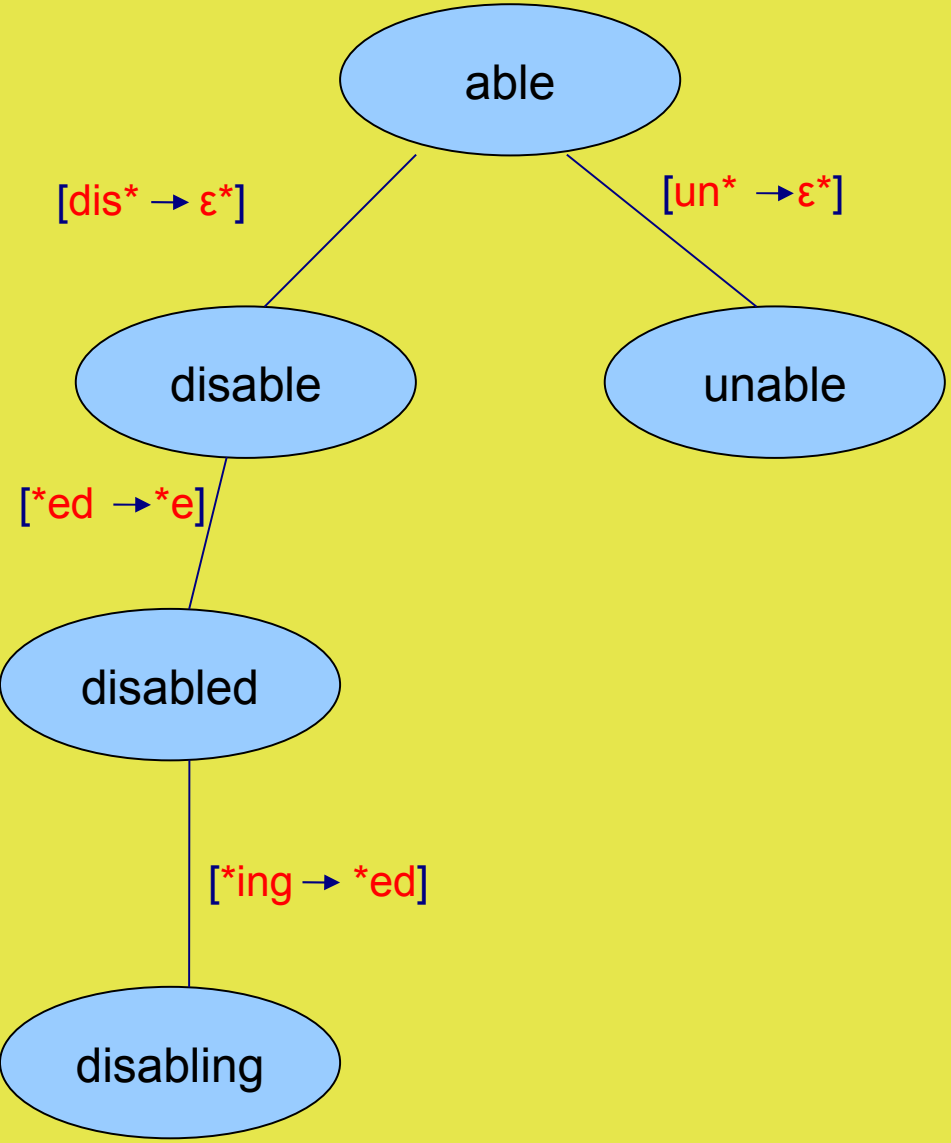
Word Relation Tree



Word Relation Tree



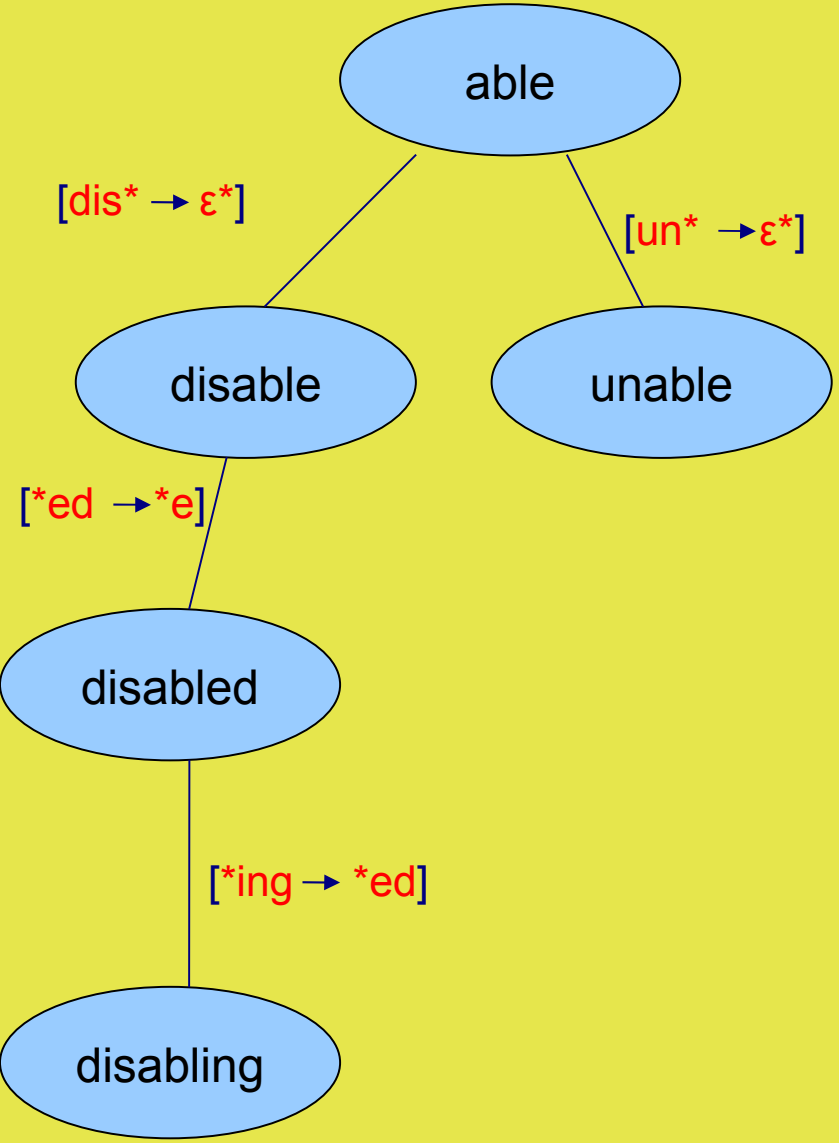
Word Relation Tree



Regroups word sharing the same stem

- The root is the stem
- The complexity increase with the depth.

Morpheme Extraction



We extract all the C-Rules linking the word to analyse and the root

Ex: disabling, able : {[**dis*** ⟶ **ε***],
[***ing** ⟶ ***ed**],
[***abled** ⟶ ***able**]}

Morpheme Extraction

The morphemes are the left part of the most frequent ***equivalent*** C-Rule.

A C-Rule is equivalent to another C-Rule if it always gives the same result.

Ex: Equivalent C-Rule of **[*abled → *able]**: **{[*abled → *able]**
[*bled → *ble],
[*led → *le],
[*ed → *e],
[*d → *ε]}

Morpheme Extraction

{[dis* → ε*], [*ing → *ed], [*abled → *able]} ⇒ {dis, ed, ing}

Morphemes found at the right part of an equivalent C-Rule are removed.

Ex: {**dis**, **ing**}

The stem is added to the list of morphemes.

Ex: {**dis**, **ing**, **able**}

Metric

Frequency is not the best metric (bias toward short C-Rules).

EX: [**anti-*** → **ε***] 2 472
[**ka*** → **ε***] 13 839

Use productivity instead

EX: [**anti-*** → **ε***] 0.949
[**ka*** → **ε***] 0.247

Results

	RALI-ANA			RALI-COF			Morfessor Baseline	Overall best system
	Pr.	Rc.	F1	Pr.	Rc.	F1	F1	Rank
ENG.	64.61	33.48	44.10	68.32	46.45	55.30	59.84	4/12
GER.	61.39	15.34	24.55	67.53	34.38	45.57	35.87	4/12
FIN.	60.06	10.33	17.63	74.76	26.20	38.81	26.75	4/11
TUR.	69.52	12.85	21.69	48.43	44.54	46.40	29.67	2/12
Arb. (V)	91.30	2.83	5.49	95.09	1.50	2.95	9.28	10/10

Results

	RALI-ANA			RALI-COF			Morfessor Baseline	Overall best system
	Pr.	Rc.	F1	Pr.	Rc.	F1	F1	Rank
ENG.	64.61	33.48	44.10	68.32	46.45	55.30	59.84	4/12
GER.	61.39	15.34	24.55	67.53	34.38	45.57	35.87	4/12
FIN.	60.06	10.33	17.63	74.76	26.20	38.81	26.75	4/11
TUR.	69.52	12.85	21.69	48.43	44.54	46.40	29.67	2/12
Arb. (V)	91.30	2.83	5.49	95.09	1.50	2.95	9.28	10/10

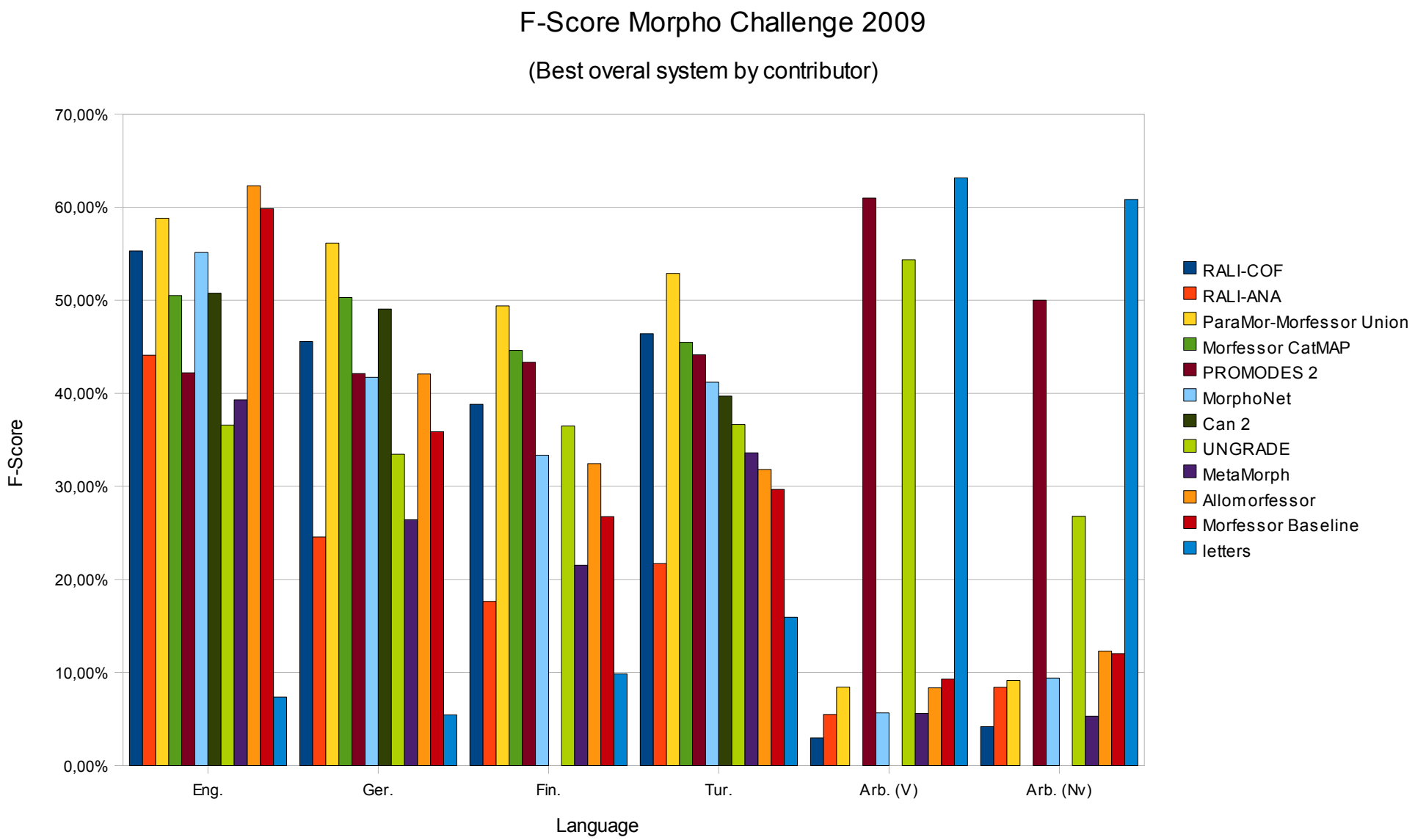
Results

	RALI-ANA			RALI-COF			Morfessor Baseline	Overall best system
	Pr.	Rc.	F1	Pr.	Rc.	F1	F1	Rank
ENG.	64.61	33.48	44.10	68.32	46.45	55.30	59.84	4/12
GER.	61.39	15.34	24.55	67.53	34.38	45.57	35.87	4/12
FIN.	60.06	10.33	17.63	74.76	26.20	38.81	26.75	4/11
TUR.	69.52	12.85	21.69	48.43	44.54	46.40	29.67	2/12
Arb. (V)	91.30	2.83	5.49	95.09	1.50	2.95	9.28	10/10

Results

	RALI-ANA			RALI-COF			Morfessor Baseline	Overall best system
	Pr.	Rc.	F1	Pr.	Rc.	F1	F1	Rank
ENG.	64.61	33.48	44.10	68.32	46.45	55.30	59.84	4/12
GER.	61.39	15.34	24.55	67.53	34.38	45.57	35.87	4/12
FIN.	60.06	10.33	17.63	74.76	26.20	38.81	26.75	4/11
TUR.	69.52	12.85	21.69	48.43	44.54	46.40	29.67	2/12
Arb. (V)	91.30	2.83	5.49	95.09	1.50	2.95	9.28	10/10

Results


F-Score Morpho Challenge 2009
(Best overal system by contributor)
70,00%
60,00%
50,00%
40,00%
30,00%
20,00%
10,00%
0,00%
F-Score
Eng.
Ger.
Fin.
Tur.
Arb. (V)
Arb. (Nv)
Language
RALI-COF
RALI-ANA
ParaMor-Morfessor Union
Morfessor CatMAP
PROMODES 2
MorphoNet
Can 2
UNGRADE
MetaMorph
Allomorfessor
Morfessor Baseline
letters

Analysis

RALI-ANA

- Good result in English
- Lower recall caused by missing analysis
- Good precision

RALI-COF

- Higher F-Score than RALI-ANA
- Good results on 4 out of 5 langages
- Could increase performance by adjusting meta-parameters per language
- Results demonstrate that the approach is viable

Analysis

Arabic

- Low F-Score
- Precision > 90%, recall < 10%
- Small training set
- Lots of morphemes for short words

ex : >atawAo:'ty faEala 'ataw +Verb +Perf +Act +3P +Pl

7 symbols 8 morphemes

Thank You

Questions ?