Minimal supervision for
(1) POS tagging
(2) gender induction

Cucerzan & Yarowsky, 2002, 2003

Summarized by Tomáš Sourada,

Language Technologies in Practice (NPFL128),

some summarizing ideas and pictures from Christian Cayralat and Jirka Hana

Part 1 (POS tagger): Bootstrapping a Multilingual Part-of-speech Tagger in One Person-day

[Cucerzan & Yarowsky, 2002]

The Research Question: how to

- build a fine-grained POS tagger
- for a low resource language
- without a native speaker of that language
- minimizing the number of person-hours invested
- ?

- recall: what is a POS tagger?
- fine-grained: *destruí -> V-pret-1sg*

Minimal Supervision - definition

- previous work:
  - only partially tagged corpora
  - small tagged seed wordlists
  - automatic transfer of annotations from another language
- this work:
  - minimal amount of person-hours needed to create the annotations
  - minimal cost needed to pay the people

Working example

- building a POS tagger for Romanian (here: low-resource language)
- use the knowledge of English (high-resource language)
- transfer the knowledge to Romanian (generally any language)

Data resources

Romanian -> English

1. Bilingual Dictionary

Romanian	True POS	English translation list
mandat	N	warrant; proxy; mandate; money order; power of attorney
manechin	N	model, dummy
manifesta	V	arise, express itself, show
manual	Adj	manual;
	N	manual; textbook; handbook
mare	Adj	large; big; great; tall; old; important;
	N	sea
maro	Adj	brown, chestnut

Figure 1: A sample Romanian-English dictionary. The POS tags are used only for evaluation and are not available in many bilingual dictionaries.

We need to find those

Romanian

2. Reference Grammar

Root Affix	Inflected Affix	Part-of-speech Tag
Spanish:		
o$	o$	Adj-masc-sing
o$	os$	Adj-masc-plur
o$	a$	Adj-fem-sing
o$	as$	Adj-fem-plur
e$	e$	Adj-masc,fem-sing
e$	es$	Adj-masc,fem-plur
ar$	o$	Verb-Indic_Pres-p1-sing
ar$	as$	Verb-Indic_Pres-p2-sing
ar$	a$	Verb-Indic_Pres-p3-sing
ar$	amos$	Verb-Indic_Pres-p1-plur
ar$	áis$	Verb-Indic_Pres-p2-plur
ar$	an$	Verb-Indic_Pres-p3-plur
Romanian:		
ă$	e$	Noun-Nomin-p3-fem-plur-indef
e$	i$	Noun-Nomin-p3-fem-plur-indef
ea$	ele$	Noun-Nomin-p3-fem-plur-indef
i$	ile$	Noun-Nomin-p3-fem-plur-indef
a$	ale$	Noun-Nomin-p3-fem-plur-indef
$	$	Adj-masc,neut-sing
$	ă$	Adj-fem-sing
$	i$	Adj-masc,neut,fem-plur
$	e$	Adj-fem,neut-plur
ru$	ra$	Adj-fem-sing
ru$	ri$	Adj-masc,neut,fem-plur
ru$	re$	Adj-fem-plur
...	...	...
e$	$	Verb-Indic_Pres-p1-sing
e$	i$	Verb-Indic_Pres-p2-sing
e$	e$	Verb-Indic_Pres-p3-sing
e$	em$	Verb-Indic_Pres-p1-plur
e$	eţi$	Verb-Indic_Pres-p2-plur
e$	$	Verb-Indic_Pres-p3-plur

Table 2: Sample extracted regular inflectional paradigms (suffix context is marked by $).

Romanian

3. Monolingual (low-resource language) unannotated corpus

Tóte oraşele unuĭ districtŭ formézǎ unŭ singurŭ colegiŭ cu oraşulŭ de reşed'nţǎ.

Art. 7. Facŭ parte din colegiulŭ al patrulea toţĭ aceia carĭ plătescŭ o dare cǎtre Statŭ orĭ-câtŭ de micǎ şi care nu intrǎ în nicĭ una di ı categoriile de maĭ susŭ.

Preoţiĭ carĭ nu arŭ face parte din nicĭ unulŭ din*colegiurile de maĭ susŭ, facŭ parte din acestŭ alŭ patrulea colegiŭ.

Acestŭ colegiŭ alege la alŭ duoilea gradŭ unŭ deputatŭ de districtŭ.

Cincĭ-decĭ de alegǎtorĭ înscrişĭ numescŭ unŭ delegatŭ.

Delegaţiĭ întruniţi la reşedinţa districtuluĭ, alegŭ pe dep"tatŭ.

Art. 8. Pentru Senatŭ, corpulŭ electoralŭ este împǎrţitŭ în fie-care judeţŭ în două colegiurĭ.

Art. 9. Primulŭ colegiŭ se compune din toţĭ proprietariĭ de fondurĭ rurale din judeţŭ, carĭ aŭ unŭ venitŭ fonciarŭ de treĭ sute galbenĭ celŭ pucinŭ.

Art. 10. Celŭ de alŭ duoilea colegiŭ se compune din toţĭ proprietariĭ de nemişcǎtóre aĭ oraşelorŭ din districtŭ, carĭ aŭ unŭ venitŭ fonciarŭ de treĭ sute galbenĭ în josŭ, potrivitŭ art. 11.

Art. 11. În oraşele unde nu s'ar gǎsi unŭ numěrŭ de una sutǎ alegětorĭ pentru a forma celŭ de alŭ duoilea colegiŭ, acestŭ numěrŭ se va complecta cu proprietariĭ judeţuluĭ, posedândŭ unŭ venitŭ fonciarŭ între treĭ sute şi una sutǎ galbenĭ, preferindu-se pururea ceĭ maĭ greŭ impuşĭ, şi orǎşaniĭ asupra proprietarilorŭ de moşiĭ.

Guideline. The task: annotate a corpus with POS tags

1. Induce Candidate POS tags:
   - token -> possible POS tags?
   - bilingual dict + English annotations -> (Rom.) POS tag distribution
2. Fine-grain it
   - *destruí: VERB -> V-pret-1sg*
   - manually extract regular rules from a (Romanian) grammar
   - improve it to match also semi-regularities and irregularities
   - manually list irregular closed-class words
3. Make it robust
   - suffix trie to deal with non-covered words
   - use monolingual corpus -> $P(pos_2|pos_1, pos_0)$, $P(w_i|pos_j)$
   - n-grams with backoff to simpler tagsets (POS only)
   - iterative re-estimation
   - gender induction (we will see)

Amount of supervision

~ 3 hour for dict extraction

~ 2 hours

~ 3 hours

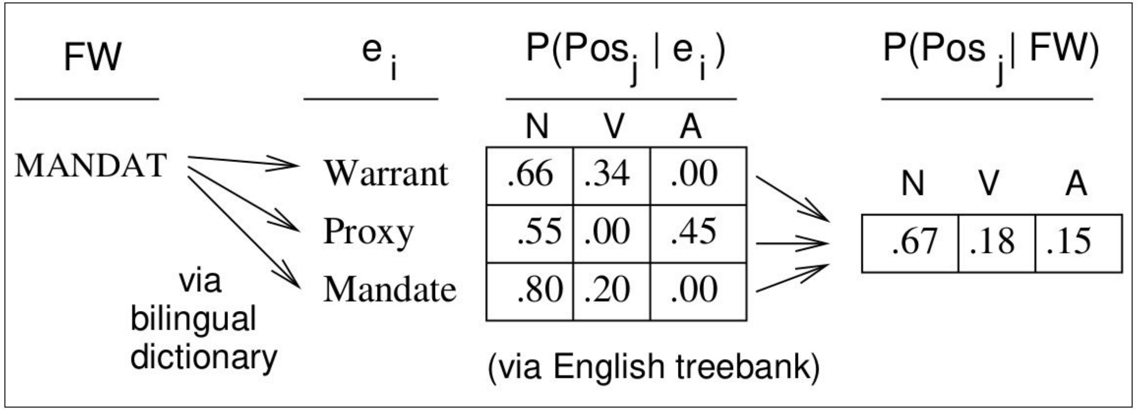
Sum: 8h (1 person-day)

1. Induce Candidate POS tags

- knowledge of POS in English + Romanian-English dictionary
  - gives candidate POS tags
- simple for words, phrases must be interpolated

$$
\begin{aligned}
P(N_f|money\ \ order) = \\
& P(N_f|N_eN_e)\cdot P(N_e|money)\cdot P(N_e|order)+\\
& P(N_f|N_eV_e)\cdot P(N_e|money)\cdot P(V_e|order)+\\
& P(N_f|V_eN_e)\cdot P(V_e|money)\cdot P(N_e|order)+\\
& P(N_f|V_eV_e)\cdot P(V_e|money)\cdot P(V_e|order)+\\
& \ldots
\end{aligned}
$$

$$
\begin{aligned}
P(T_f|w_{e_1}...w_{e_n}) = \\
& P(T_f|T_{e_1}...T_{e_n})\cdot P(T_{e_1}...T_{e_n}|w_{e_1}...w_{e_n})
\end{aligned}
$$



1. Induction Results

POS with highest predicted probability is taken

treshold prob = 0.1

percentage of words for whose at least something was predicted

probability mass associated with the true POS tag averaged over all words

Target Language	Training Dictionary	Accuracy Exact POS	Correct POS Over Threshold	Coverage	Mean Probability of Truth
Romanian	Spanish - English	92.9	97.8	98	.91
Kurdish	Spanish - English	76.8	93.1	95	.82
Spanish	Romanian - English	83.3	94.9	97	.86

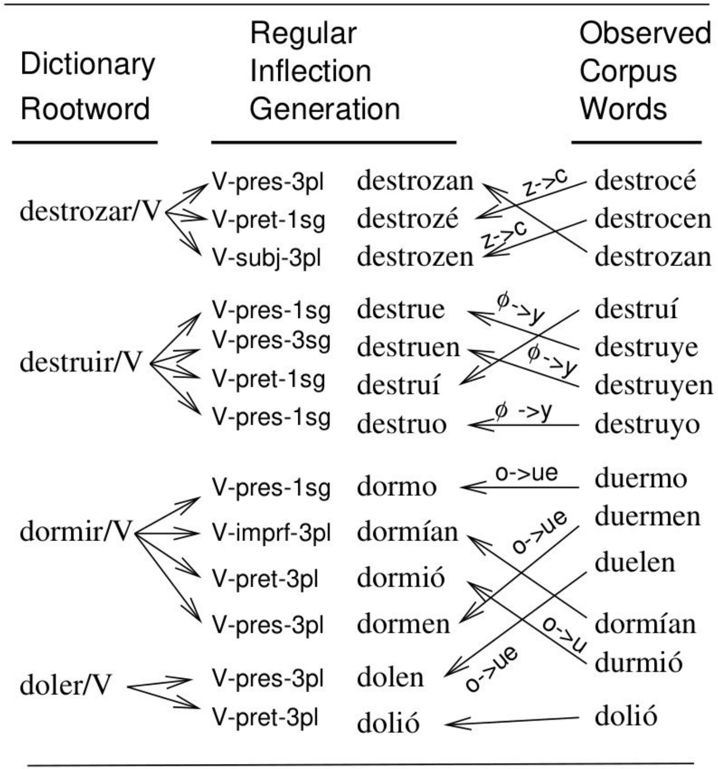
Table 1: Performance of inducing candidate part-of-speech distributions derived solely from untagged English translation lists. Results are measured by type (all dictionary entries are weighted equally).

2. Fine-graining through morphological analysis

2A. Manually extract:

Root Affix	Inflected Affix	Part-of-speech Tag
Spanish:		
o$	o$	Adj-masc-sing
o$	os$	Adj-masc-plur
o$	a$	Adj-fem-sing
o$	as$	Adj-fem-plur
e$	e$	Adj-masc,fem-sing
e$	es$	Adj-masc,fem-plur
ar$	o$	Verb-Indic_Pres-p1-sing
ar$	as$	Verb-Indic_Pres-p2-sing
ar$	a$	Verb-Indic_Pres-p3-sing
ar$	amos$	Verb-Indic_Pres-p1-plur
ar$	áis$	Verb-Indic_Pres-p2-plur
ar$	an$	Verb-Indic_Pres-p3-plur

2B. Improve using Levenshtein alignment:



2. Fine-graining through morphological analysis

2C: manually list closed-class words

- with their fine-grained tags
- *ser, mi, tu, su, aquel*

3. Make it robust

- 3A: suffix trie to increase coverage to unseen words
- 3B: n-grams with back-off to simpler tagsets (part-of-speech only)
- 3C: iterative re-estimation
- (gender: the other paper)

Results

- a lot of errors due to inconsistent annotation
- in Romanian, additional 4 hours of native speaker work for comparison
- good results both with core-tags and fine-grained tags
- 1 person-day suffices
  - (compare with $100,000-$1,000,000
  - spent on annotating corpora)

	Spanish	Romanian	
	NNS 8h	NNS 8h	NNS-8h NS-4h
All words			
core-tag	93.1	86.3	89.2
exact-match	86.5	68.6	75.5
exact w/o gender	87.0	76.7	83.0

Conclusion

- we can get a POS tagger
- after 1 person-day of work
- for any language that has
  - reference grammar
  - bilingual dictionary (to English)
  - large enough monolingual corpus (megawords used)

Part 2 (gender): Minimally Supervised Induction of Grammatical Gender

[Cucerzan & Yarowsky, 2003]

Induce grammatical gender (masculine, feminine, neuter)

- Motivation:
  - knowing gender is important in POS tagging
  - can be important in NLG systems, MT systems (noun-adjective agreement etc.)
- previous work:
  - POS taggers induced gender during prediction
  - important (difficult) only for nouns, for the rest it is easy by agreement
- this work:
  - induce gender independently of other task
  - language-independent approach (well, not really)
  - minimal supervision required

Recall

- what is precision and what is coverage (aka recall)?

The approach

1. seeds
   - ~50 seed nouns with known gender (need of supervision, high precision (100%), extremely low coverage (~0.1%))
2. bootstrapping using context
   - seeds -> contexts that determine the gender -> more nouns with reliable gender
   - iterate
   - still high precision (~99%), still low coverage (~50%)
3. morphological model
   - based on suffix-similarity predict gender of most of the rest
   - lower precision (~98%), high coverage (almost 100%)
4. dealing with special cases
   - words with rare endings, do not share suffix with any other word
   - predict the class (gender) with the most variability of suffixes

1. Seeding: how to obtain ~50 nouns with gender annotation?

Method 1 - Translingual Projection of Natural Gender

- in English, we know the natural gender of some nouns
- translate them to obtain the seed nouns in a new language
- need to remove colliding translations
- limitation: ? collision of grammatical and natural gender

Feminine	Freq	R/F/E/S	Masculine	Freq	R/F/E/S
woman	322	+/+/+/+	man	1396	+/+/+/+
girl	234	$\stackrel{+}{=}$/+/$\stackrel{+}{=}$/+	boy	261	$\stackrel{+}{=}$/+/$\stackrel{+}{=}$/+
sister	56	+/+/+/+	brother	106	+/+/+/+
mother	268	+/+/+/+	father	246	+/+/+/+
wife	302	+/+/+/+	husband	184	+/+/+/+
daughter	93	$\stackrel{+}{=}$/+/+/+	son	191	$\stackrel{+}{=}$/+/+/+
daughter-in-law	1	+/+/+/*	son-in-law	5	+/+/+/*
stepdaugther	1	?/?/+/+	stepson	3	?/?/+/+
grandmother	14	?/+/+/+	grandfather	17	?/+/+/*
granddaughter	3	+/+/+/+	grandson	7	+/+/+/+
aunt	11	+/+/?/+	uncle	26	+/+/+/+
niece	9	+/+/+/+	nephew	11	+/+/+/+
bride	39	?/+/+/+	groom	5	?/?/+/+
girlfriend	5	?/?/$\stackrel{+}{=}$/?	boyfriend	1	+/?/$\stackrel{+}{=}$/?
lady	62	+/?/+/+	gentleman	26	+/?/+/+
mistress	8	?/+/+/+	mister	5	?/?/?/+
queen	26	+/+/±/+	king	42	+/+/+/+
princess	7	?/+/+/+	prince	6	+/+/+/+
governess	4	+/?/+/*	governor	84	?/+/±/+
duchess	1	?/+/+/*	duke	6	+/+/+/+
empress	0	?/+/+/+	emperor	11	+/+/?/+
baroness	2	?/+/+/+	baron	3	?/+/+/+
witch	10	?/+/+/*	soldier	43	+/+/+/+
actress	17	+/+/$\stackrel{+}{=}$/+	actor	43	+/+/$\stackrel{+}{=}$/+
waitress	4	+/+/$\stackrel{+}{=}$/+	waiter	11	+/+/$\stackrel{+}{=}$/+
mare	15	+/?/+/+	stallion	7	+/?/+/*
cow	30	+/+/+/+	bull	29	+/+/±/+
bitch	8	+/+/+/*	dog	85	+/+/+/+
hen	23	+/±/?/+	rooster	5	?/+/+/?
doe	1	?/?/+/*	stag	9	+/?/+/+
	1575			2874	

1. Seeding: how to obtain ~50 nouns with gender annotation?

Method 2 - Frequency-based extraction:

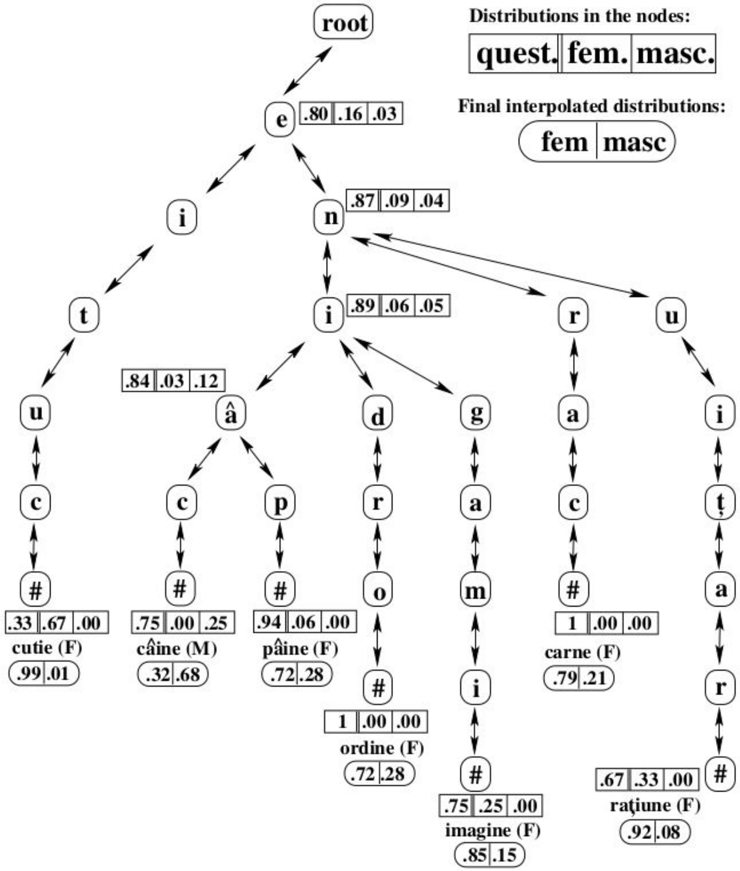
- Extract nouns from corpus on the basis of:
    - frequency
    - number of contexts (gender agreement) with which they occur
    - suffix patterns
- manually label gender -> need of gender-annotated dictionary
- guarantees representativeness of the seeds
- unclear description of HOW they did it (what does “extraction on the basis of frequency, …” mean)

2. Bootstrapping using context

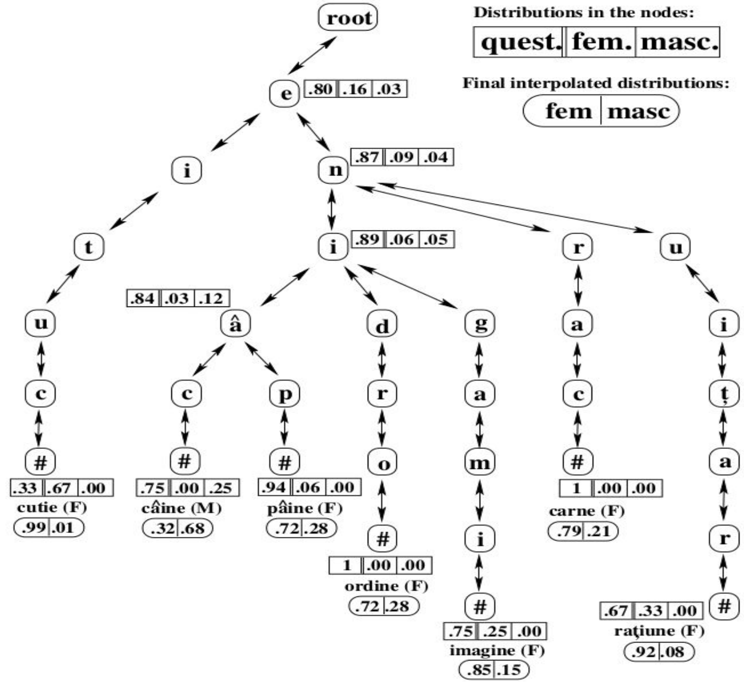
- 6 different contexts: {left, right, bilateral} x {whole words, word suffixes}
- unclear: what are suffixes? (word endings)
- main method:
  - select contexts that occur a lot with the seed nouns
  - if the gender of the context can be determined reliably (over a treshold), mark the context with the gender
  - add new nouns to the seed list (those that appear mostly in the context)
  - iterate
- -> high precision (~99%), low coverage (~50%)
- assumption: the gender of a word is reflected in the context

3. Morphology models: suffix-based induction of gender

- language dependent!
- words with long common ending (here =suffix) usually share the gender
- weighted combination of words with the longest common suffix and words with shorter (yet longer than 0) common suffix
- suffix trie used for effective implementation



3. Morphology models: suffix-based induction of gender



$$\lambda_{node,\alpha,\beta} = \frac{1-\beta P_{node(l_{n:i})}(quest)^{\alpha}}{1-P_{node(l_{n:i})}(quest)}$$

β: how much prob. mass to transfer from node to node

$\alpha \in (0,\infty), \beta \in (0,1)$

$$\widehat{P}(gen_j|l_n l_{n-1}...l_i) = P_{node(l_n l_{n-1}...l_i)}(gen_j) + P_{node(l_n l_{n-1}...l_i)}(quest) \cdot \widehat{P}(gen_j|l_n l_{n-1}...l_{i+1})$$

Interpolation

$$\widehat{P}(gen_j|l_n l_{n-1}...l_i) = \lambda_{node,\alpha,\beta} P_{node(l_n l_{n-1}...l_i)}(gen_j) + \beta P_{node(l_n l_{n-1}...l_i)}(quest)^{\alpha} \cdot \widehat{P}(gen_j|l_n l_{n-1}...l_{i+1})$$

Recursion

4. Dealing with special cases

- there are words whose gender cannot be induced even by the morphology model
  - words with weird endings, unseen characters
  - e.g. single letters *A, B, C*
- two options:
  - predict the most likely (frequent) class (M.L.)
  - predict the class with the most variable endings (M.V.) - empirically better

	Romanian	French	Spanish	Slovene	Swedish
unk	0.19%	0.08%	0.03%	0.46%	0.09%
M.L.	0	0	100	10.00	41.18
M.V.	100	100	100	90.00	41.18

Table 4: Percentage of nouns for which predictions cannot be made and the accuracy obtained for these nouns by predicting the most likely class (M.L.) and the class with most endings (M.V.) in the language

Possible improvements

- Problem: low coverage after context bootstrapping
  - precision-recall tradeoff
  - caused by limited size of used corpus
  - superior results when using web search (be aware, it was 2002, but still really large corpus)
    - 100% accuracy, 94% coverage

Results - French, Spanish

- 2 types of evaluation:
  - by type (all nouns treated as equally important)
  - by token (weighted by type frequency)
- coverage vs. accuracy

French	Natural gender seeds (31 fem., 35 masc.)			
	by type		by token	
1317 nouns	context	+morph.	context	+morph.
coverage	77.15	100	86.00	100
accuracy	**97.51**	**95.44**	98.26	97.18
French	System extracted seeds (19 fem., 29 masc.)			
	by type		by token	
1317 nouns	context	+morph.	context	+morph.
coverage	76.31	100	94.28	100
accuracy	**99.50**	**96.81**	99.73	98.81

Table 6: Results for French

Spanish	Natural gender seeds (53 fem., 51 masc.)			
	by type		by token	
2993 nouns	context	+morph.	context	+morph.
coverage	54.06	100	72.71	100
accuracy	**98.70**	**95.59**	99.47	98.45
Spanish	System extracted seeds (18 fem., 30 masc.)			
	by type		by token	
2993 nouns	context	+morph.	context	+morph.
coverage	50.84	100	77.33	100
accuracy	**98.69**	**95.49**	99.51	98.13

Table 7: Results for Spanish

Results - Slovene and Swedish

Slovene	Natural gender seeds (44 fem., 40 masc.)			
	by type		by token	
2170 nouns	context	+morph.	context	+morph.
coverage	2.26	100	3.64	100
accuracy	**100**	**90.60**	100	78.32

Slovene	System extracted seeds (27 fem., 19 masc.)			
	by type		by token	
2170 nouns	context	+morph.	context	+morph.
coverage	18.99	100	64.86	100
accuracy	**99.51**	**95.62**	98.18	96.71

Table 8: Results for Slovene

Swedish	Natural gender seeds (38 ~fem., 41 ~masc.)			
	by type		by token	
19877 nouns	context	+morph.	context	+morph.
coverage	0.30	100	1.81	100
accuracy	44.07	46.21	46.21	45.92

Swedish	System extracted seeds (27 comm., 23 neut.)			
	by type		by token	
19877 nouns	context	+morph.	context	+morph.
coverage	35.61	100	72.73	100
accuracy	**98.84**	**94.41**	99.62	96.50

Table 9: Results for Swedish

- results in Swedish, natural gender seeds is close to random
    - because Swedish gender does not follow standard feminine/masculine distinction

Thanks for your attention.