

# Machine Learning for NLP

Learning from small data: low resource languages

---

Aurélie Herbelot

2018

Centre for Mind/Brain Sciences  
University of Trento

- What are low-resource languages?
- High-level issues.
- Getting data.
- Projection-based techniques.
- Resourceless NLP.

What is 'low-resource'?

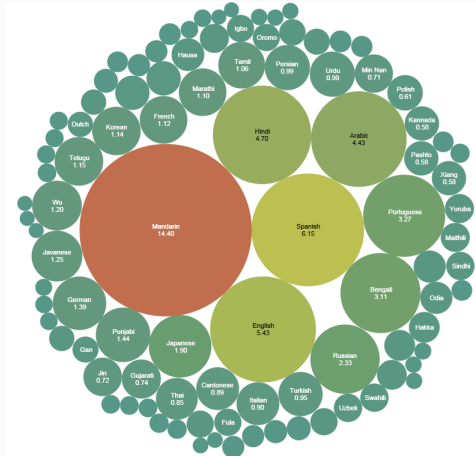
# Languages of the world

**Table 2. Distribution of world languages by number of first-language speakers**

Population range	Living languages			Number of speakers		
	Count	Percent	Cumulative	Total	Percent	Cumulative
100,000,000 to 999,999,999	8	0.1	0.1%	2,709,546,730	40.78777	40.78777%
10,000,000 to 99,999,999	82	1.2	1.3%	2,609,446,190	39.28092	80.06869%
1,000,000 to 9,999,999	307	4.3	5.6%	948,917,508	14.28439	94.35308%
100,000 to 999,999	956	13.5	19.1%	305,209,791	4.59443	98.94751%
10,000 to 99,999	1,811	25.5	44.6%	61,803,881	0.93036	99.87787%
1,000 to 9,999	1,980	27.9	72.5%	7,630,091	0.11486	99.99272%
100 to 999	1,064	15.0	87.4%	470,472	0.00708	99.99981%
10 to 99	329	4.6	92.1%	12,268	0.00018	99.99999%
1 to 9	144	2.0	94.1%	584	0.00001	100.00000%
0	219	3.1	97.2%	0	0.00000	100.00000%
Unknown	199	2.8	100.0%			
<i>Totals</i>	7,099	100.0		6,643,037,515	100.00000	

<https://www.ethnologue.com/statistics/size>

# Languages of the world

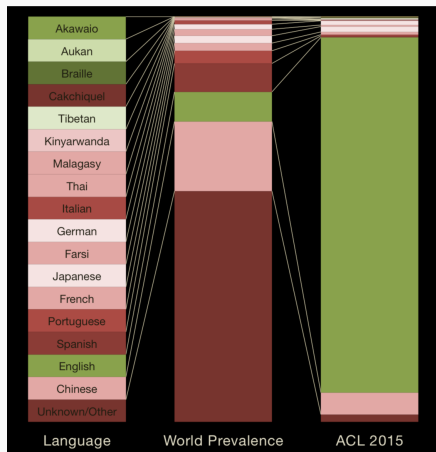


Languages by proportion of native speakers,

<https://commons.wikimedia.org/w/index.php?curid=41715483>

# NLP for the languages of the world

- The ACL is the most prestigious computational linguistic conference, reporting on the latest developments in the field.
- How does it cater for the languages of the world?



<http://www.junglelightspeed.com/languages-at-acl-this-year/>

## NLP research and low-resource languages (Robert Munro)

- 'Most advances in NLP are by 2-3%.'
- 'Most advantages of 2-3% are specific to the problem and language at hand, so they do not carry over.'
- 'In order to understand how computational linguistics applies to the full breath of human communication, we need to test the technology across a representative diversity of languages.'
- 'For vocabulary, word-order, morphology, standardized of spelling, and more, English is an outlier, telling little about how well a result applies to the 95% of the worlds communications that are in other languages.'

# The case of Malayalam



സ്വതന്ത്ര  
മലയാളം  
കമ്പ്യൂട്ടിംഗ്

## Swathanthra Malayalam Computing

Swathanthra Malayalam Computing (SMC) is a free software collective engaged in development, localization, standardization and popularization of various Free and Open Source Softwares in Malayalam language. "എന്റെ കമ്പ്യൂട്ടറിന് എന്റെ ഭാഷ" is the slogan of the organization, which translates to "My language for/on My Computer".

- Malayalam: 38 million native speakers.
- Limited resources for font display.
- No morphological analyser (extremely agglutinative language), POS tagger, parser...
- Solutions for English do not transfer to Malayalam.



## A case in point: automatic translation

- The back-and-forth translation game...
- Translate sentence  $S_1$  from language  $L_1$  to language  $L_2$  via system  $T$ .
- Use  $T$  to translate  $S_2$  back into language  $L_1$ .
- Expectation:  $T(S_1) = S_2$  and  $T(S_2) \approx S_1$ .

# Google translate: English <=> Malayalam

Malayalam English Nepali Detect language ▾

↔

Nepali English Malayalam ▾

Translate

The dog plays with the ball. | ×

28/5000

പന്ത് പന്ത് കൊണ്ട് കളിക്കുന്നു. ×

32/5000

☆ 📄 🔊 ↻

Suggest an edit

☆ 📄 🔊 ↻

# Google translate: English <→> Chichewa

Malayalam	English	Chichewa	Detect language			Chichewa	English	Malayalam		Translate
<p>The dog plays with the ball.</p> <p>28/5000</p>					×	<p>The galu amaseweretsa mpira.</p> <p>☆ 📄 ↶</p>				
<p>The galu amaseweretsa mpira.</p> <p>28/5000</p>					×	<p>The dog's toy ball.</p> <p>☆ 📄 🔊 ↶</p>				

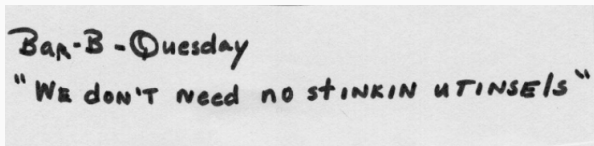
# High-level issues in processing low-resource languages

# Language documentation and description

- The task of collecting samples of the language (traditionally done by field linguists).
- A lot of the work done by field linguists is unpublished or in paper form! Raw data may be hard to obtain in digitised format.
- For languages with Internet users, the Web can be used as a (small) source of raw text.
- Bible translations are often used! (Bias issue...)
- Many languages are primarily oral.

## Pre-processing: orthography

- Orthography for a low-resource language may not be standardised.
- Non-standard orthography can be found in any language, but some lack standardisation entirely.
- Variations can express cultural aspects.

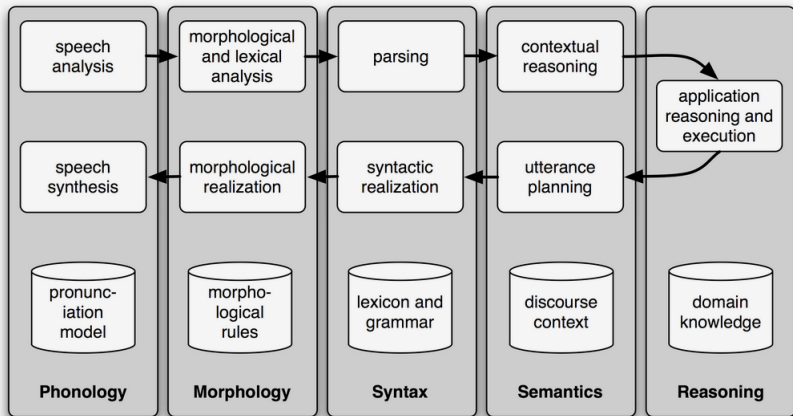


Alexandra Jaffe. Journal of sociolinguistics 4/4. 2000.

# What is a language?

- Does the data belong to the same language?
- As long as mutual intelligibility has been shown, two seemingly different data sources can be classed as dialectal variants of the same language.
- The data may exhibit complex variations as a result.

# The NLP pipeline



Example NLP pipeline for a Spoken Dialogue System.

[http://www.nltk.org/book\\_1ed/ch01.html](http://www.nltk.org/book_1ed/ch01.html).



## Gathering data

# A simple Web-based algorithm

- Goal: find Web documents in a target language.
- Crawling the entire Web and classifying each document separately is clearly inefficient.
- *The Crúbadán Project* (Scannell, 2007): use search engines to find appropriate documents:
  - build a query of random words of the language, separated by *OR*
  - append one frequent (and unambiguous) function word from that language.

```
agus AND sainchomhairle OR ndamhsa OR oirfidigh OR caillteacha OR rancás
```

## Encoding issues: examples

- **Mongolian:** most Web documents are encoded as CP-1251.
- In CP-1251, decimal byte values 170, 175, 186, and 191 correspond to Unicode U+0404, U+0407, U+0454, and U+0457.
- In Mongolian, those bytes are supposed to represent U+04E8, U+04AE, U+04E9, and U+04AF... (Users have a dedicated Mongolian font installed.)
- **Irish:** before 8-bit email, users wrote acute accents using ‘/’: *be/al* for *béal*.
- Because of this, the largest single collection of Irish texts (on *listserve.heatnet.ie*) is invisible through Google (which treats ‘/’ as a space).

- Google retired its API a long time ago...
- There is currently no easy way to do (free) intensive searches on (a large proportion of) the Web.

# Language identification

- How to check that the retrieved documents are definitely in the correct language?
- Performance on language identification is quite high (around 99%) when enough data is available.
- It however decreases when:
  - classification must be performed over many languages;
  - texts are short.
- Accuracy on Twitter data is less than 90% (1 error in 10!)

# Multilingual content

## stultus@web:~/Blog\$

My life. My choices. My problems. My mistakes. My lessons.


Log in

Categories: Audio Books Debian Freedom Fun Hack Hactivism Input International Chalu Union InternetFreedom keyboard Knowledge LaTeX Life Linux NetNeutrality People Published Elsewhere quote smc Tv Discussion Video wikipedia കവിത മലയാളം

Search

### ഭോളുകളുടെ ജനാധിപത്യം - മനോരമ ന്യൂസിലെ നിയന്ത്രണരേഖയിൽ

മനോരമ ന്യൂസിലെ നിയന്ത്രണരേഖയിൽ ഭോളുകളുടെ ജനാധിപത്യം എന്ന വിഷയത്തിൽ നടന്ന ചർച്ചയിൽ ഐ.സി.ഡു. വിനെ പ്രതിരോധിച്ച് പങ്കെടുത്തു. വി.ടി ബൽറാം, ഉഴപ്പൂർ വിജയൻ, വി.വി രാജേഷ്, ആർദ്ര നന്യാർ, സുജാത് നായർ എന്നിവരും പങ്കെടുത്തിരുന്നു



#### SUBSCRIBE

Site RSS Feed

#### About stultus@web:~/Blog\$

Hrishi's experiments with freedom and life...

#### Recent Posts

- ഭോളുകളുടെ ജനാധിപത്യം - മനോരമ ന്യൂസിലെ നിയന്ത്രണരേഖയിൽ
- നെറ്റ്സ്റ്റാലിന്റെ ഉപയോഗരീതികളുടെ വിജയം
- Malayalam input in debian jessie (Gnome 3.14.4 , Ibus)
- Phone in program on AIR fm
- പ്രിബേരികൾ എന്തിന് ഈ വിഖ്യാത നായർ അണിയണമ?

#### Family

- Bela Chechi
- Jithinnetan
- Nandaja

#### Friends

- Anivar Aravind
- Aswati Jose
- Ershad K
- Haris Ibrahim K. V.
- Jishnu Mohan
- Manoj K Mohan
- Praveen Arimbrathodiyil

#### Recent Comments

- Jishnu on How to neutralize the escape key (keycode 9), without spending any XM
- Stultus on How to neutralize
- Rajesh K Nambiar
- Santhosh Thottingal
- Vasudev Kamath

- Multilingual content is common in low-resource languages.
- Speakers are often (at least) bilingual, speaking the most common majority language close to their community.
- Encoding problems, as well as linking to external content, makes it likely that several languages will be mixed.

# Code-switching

- Incorporation of elements belonging to several languages in one utterance.
- Switching can happen at the utterance, word, or even morphology level.
- “*Ich bin mega-miserably dahin gewalked.*”

NEP-EN	My car at the workshop for a much needed repairs... <i>ABA pocket khali hune bho</i> (My car at the workshop for a much needed repairs. . . now my pocket will be empty)
SPA-EN	<i>Por primera vez veo a @username actually being hateful! it was beautiful:)</i> (For the first time I get to see @username actually being hateful! it was beautiful:)

Solorio et al (2014)



## Another text classification problem...

- Language classification can be seen as a specific text classification problem.
- Basic N-gram-based methods apply:
  - Convert text into character-based N-gram features:  
 $TEXT \rightarrow \_T, TE, EX, XT, T\_ \text{ (bigrams)}$   
 $TEXT \rightarrow \_TE, TEX, EXT, XT\_ \text{ (trigrams)}$
- Convert features into frequency vectors:  
 $\{\_T : 1, TE : 1 : AR : 0, T\_ : 1\}$
- Measure vector similarity to a 'prototype vector' for each language, where each component is the probability of an N-gram in the language.

## Advantages of N-grams over lexicalised methods

- A comprehensive lexicon is not always available for the language at hand.
- For highly agglutinative languages, N-grams are more reliable than words:  
*evlerinizden* – > *ev-ler-iniz-den* → *house-plural-your-from*  
→ *from your houses* (Turkish)
- The text may be the result of an OCR process, in which case there will be word recognition errors which will be smoothed by N-grams.

## From monolingual to multilingual classification

- The Linguini system (Prager, 1999).
- A mixture model: we assume a document is a combination of languages, in different proportions.
- For a case with two languages, a document  $d$  is modelled as a vector  $k_d$  which approximates  $\alpha f_1 + (1 - \alpha)f_2$ , where  $f_1$  and  $f_2$  are the prototype vectors of languages  $L1$  and  $L2$ .

## Example mixture model

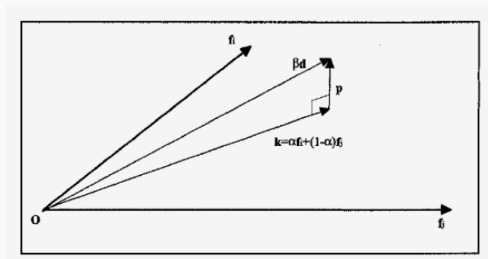
Language	Features	English Equivalent
French	le	the (masc. sing.)
	mes	my (plural)
	son	his (masc. sing)
Italian	il	the (masc. sing.)
	le	the (fem. plural)
Spanish	mes	month
	son	are (3rd person)

- Given the arbitrary ordering [il, le, mes, son], we can generate three prototype vectors:
  - French: [0,1,1,1]
  - Italian: [1,1,0,0]
  - Spanish [0,0,1,1]
- A 50/50 French/Italian model will have mixture vector [0.5, 1, 0.5, 0.5].

## Elements of the model

- A document  $d$  to classify.
- A hypothetical mixture vector  $k_d \approx \alpha f_1 + (1 - \alpha)f_2$ .
- We want to find  $k_d$  – i.e. the parameters  $(f_1, f_2, \alpha)$  – so that  $\cos(d, k_d)$  is minimum.

## Calculating $\alpha$



- There is a plane formed by  $f_1$  and  $f_2$ , and  $k_d$  lies on that plane.
- $k_d$  is the projection  $p$  of some multiple  $\beta$  of  $d$  onto that plane. (Any other vector would have a greater cosine with  $d$ .)
- So  $p = \beta d - k$  is perpendicular to the plane, and to  $f_1$  and  $f_2$ .  
 $f_1 \cdot p = f_1 \cdot (\beta d - k_d) = 0$   
 $f_2 \cdot p = f_2 \cdot (\beta d - k_d) = 0$
- From this we calculate  $\alpha$ .

## Finding $f_1$ and $f_2$

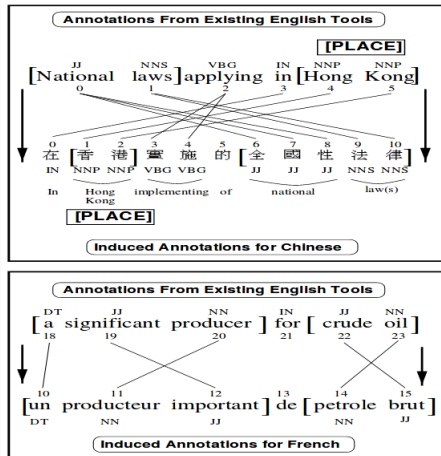
- We can employ the brute force approach and try every possible pair  $(f_1, f_2)$  until we find maximum similarity.
- Better approach: rely on the fact that if  $d$  is a mixture of  $f_1$  and  $f_2$ , it will be fairly close to both of them individually.
- In practice, the two components of the document are to be found in the 5 most similar languages.

# Projection



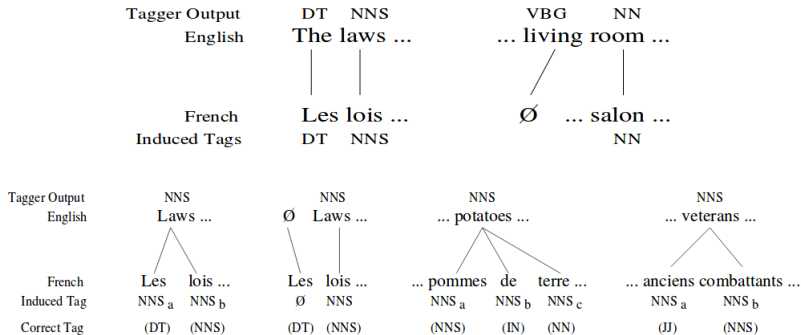
# Using alignments (Yarovsky et al, 2003)

- Can we learn a tool for a low-resource language by using one in a resourced language?
- The technique relies on having parallel text.
- We will briefly look at POS tagging, morphological induction, and parsing.



- Four-step process:
  1. Use an available tagger for the source language  $L_1$ , and tag the text.
  2. Run an alignment system from the source to the target (parallel) corpus.
  3. Transfer tags via links in the alignment.
  4. Generalise from the noisy projection to a stand-alone POS tagger for the target language  $L_2$ .

# Projection examples



## Lexical prior estimation

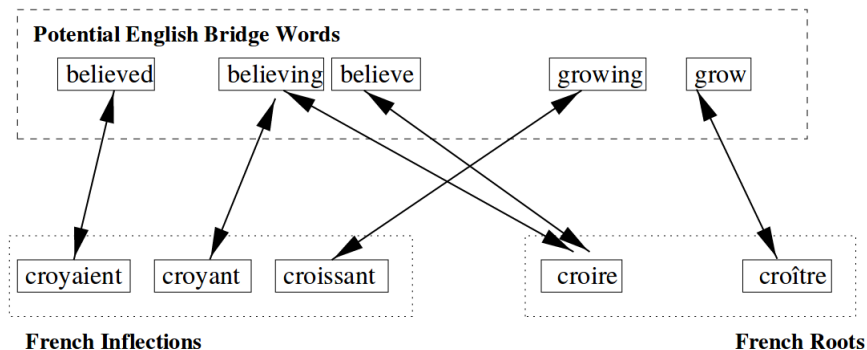
- The improved tagger is supposed to calculate  $P(t|w) \approx P(t)P(w|t)$ .
- Can we improve on the prior  $P(t)$ ?
- In some languages (French, English, Czech), there is a tendency for a word to have a high-majority POS tag, and to rarely have two.
- So we can emphasise the majority tag(s) by reducing the probability of the less frequent tags.

## Tag sequence model estimation

- We can give more or less confidence to a particular tag sequence, by estimating the quality of the alignment.
- Read out alignment score for each sentence and modify the learning algorithm accordingly.
- Most drastic solution: do not learn from alignments that score low.

# Morphological analysis induction

How can we learn that in French, *croyant* is a form of *croire*, while *croissant* is a form of *croître*?



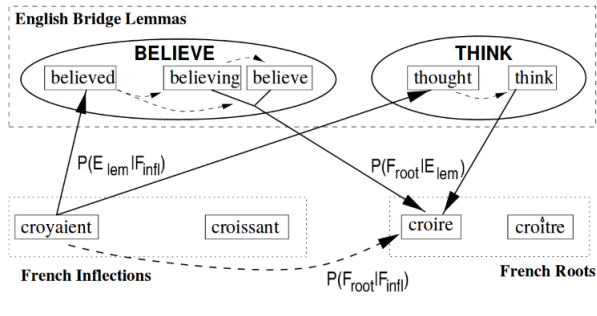
## Probability of two forms being morphologically related

- We want to calculate  $P_m(F_{root}|F_{infl})$  in the target language  $L2$ : the probability of a certain root given an inflected form.
- We assume we know clusters of related forms in  $L1$ , the source alignment language (which has an available morphological analyser).
- We build 'bridges' between the two forms via  $L1$ :

$$P_m(F_{root}|F_{infl}) = \sum_i P_a(F_{root}|F_{lem_i})P_a(F_{lem_i}|F_{infl})$$

where  $lem_i$  are clusters of word forms in  $L1$ , and  $P_a$  represents the probability of an alignment.

# Bridge alignment

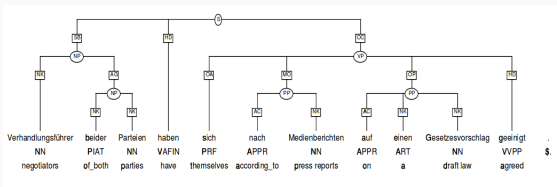


$$P_m(croire|croyaient) = P_a(croire|BELIEVE)P_a(BELIEVE|croyaient) + P_a(croire|THINK)P_a(THINK|croyaient)...$$



# Projected dependency parsing: motivation

- Learning a parser requires a treebank.
- Acquiring 20,000-40,000 sentences can take 4-7 years (Hwa et al, 2004), including:
  - building style guides
  - redundant manual annotation for quality checking.
- Not feasible for many languages!



# Projected dependency parsing (Hwa et al, 2004)

Relation $R$	Head $x_{\text{Eng}}$	Modifier $y_{\text{Eng}}$	Head $x_{\text{Bsq}}$	Modifier $y_{\text{Bsq}}$
verb-subj	got	I	erosi	nik
verb-obj	got	gift	erosi	opari
noun-det	gift	a	opari	bat
noun-mod	brother	my	anaiari	nire

## Projected dependency parsing

- We need to know that a language pair is amenable to transfer. We will have even more variability for parsing than we have for e.g. POS tagging.
- We can check this through a small human annotation of pairs of parses, over ‘perfect’ training data (i.e. manually produced parses and alignment).
- Hwa et al found a direct (unlabeled dependency) score of:
  - 38% for English - Spanish;
  - 37% for English - Chinese.

- Language-specific markers: Chinese verbs are often followed by an aspectual marker, not realised in English. This remains unattached in the projection.
- Tokenisation: Spanish clitics are separated from verbs at tokenisation stage, and produce unattached tokens:
  - *Ella va a dormirse* <--> *She's going to fall asleep*
  - After tokenisation: *Ella va a dormir se.*

# Rules-enhanced projection

- It is possible to boost the performance of the projection by adding a set of linguistically-motivated rules to the projection.
- Example: in Chinese, an aspectual marker should modify the verb to its left.
- Transformation rules: if  $f_k \dots f_n$  is followed by  $f_a$ , and  $f_a$  is an aspectual marker, make  $f_a$  modify  $f_n$ .

	Direct Projection	Projection + Transformation
English-Spanish	36.8	70.3
English-Chinese	38.1	67.3

## Additional filtering

- We can further use heuristics to filter out aligned parses that we think will be of poor quality.

- Discard if more than 20% of the English words have no Spanish counterpart.
- Discard if more than 30% of the Spanish words have no English counterpart.
- Discard if more than 4 Spanish words were aligned to the same English word.

## Real-life results

- Using manual correction rules (which took a month to write), Hwa et al's projected parser achieves a performance comparable to a commercial parser for Spanish.
- For Chinese, things are less positive...

Method	Corpus	Train Size	Parsing Performance
Baseline (mod prev)	–	–	33.8%
Stat. parser	UN/FBIS/Bible (no filter)	98K sents	67.3%
Stat. parser	UN/FBIS/Bible (w/ filter)	20K sents	72.1%
Commercial parser	–	–	69.2%

Spanish

## Real-life results

- Using manual correction rules (which took a month to write), Hwa et al's projected parser achieves a performance comparable to a commercial parser for Spanish.
- For Chinese, things are less positive...

Method	Corpus	Train Size	Parsing Performance
Baseline (mod next)	–	–	35.1%
Baseline + transformations	–	–	44.3%
Stat. parser	FBIS (w/ filter)	50K sents	53.9%
Stat. parser	ChTB (new in v4)	10K sents	64.3%

Chinese



## Delexicalised transfer parsing

- We assume access to a treebank and uses the same POS tagset as the target language.
- We train a parser on the POS tags of the source language.  
*Lexical information is ignored.*
- The trained parser is run directly onto the target language.

## The alternative: unsupervised parsing

- Since the target language is missing a treebank, unsupervised methods seem appropriate.
- A grammar can be learnt on top of POS-annotated data.
- But unsupervised parsing still lags behind supervised methods.

When there is no parallel text...

# What to do when no resource is available?

- What to do if we have:
  - no annotated corpus (and therefore no alignment);
  - no prior NLP tool – even rule-based?
- Let's see an example of POS tagging.

## Using other languages as stepping stones (Scherrer & Sagot, 2014)

- Given the target language  $L_2$ , find a language  $L_1$  which (roughly) satisfies the following:
  - $L_1$  and  $L_2$  share a lot of **cognates**: words which look similar and mean the same thing.
  - Word order is similar in both languages.
  - The set of POS tags for  $L_1$  and  $L_2$  is identical.

## The general approach

- Induce translation lexicon using a) cognate detection; b) cross-lingual context similarity.  
→  $(w_1, w_2)$  translation pairs.
- Use translation pairs to transfer POS information from  $L_1$  to  $L_2$ .
- Words still lacking a POS are tagged based on suffix analogy.

## C-SMT models

- *C-SMT* (character-level SMT) systems perform alignment at the character level rather than at the word level.
- A C-SMT model allows us to translate a word into another (presumably cognate) word.
- Generally, C-SMT models are trained on aligned data, like any SMT model.
- Without alignment available, we can try and learn the model from pairs captured with orthographic similarity measures.

# Orthographic similarity measures

- **Edit-string / Levenshtein distance:** Number of insertions/substitutions/deletions between two strings:
  - kitten  $\rightarrow$  sitten (substitution)
  - sitten  $\rightarrow$  sittin (substitution)
  - sittin  $\rightarrow$  sitting (insertion).
- **Longest Common Subsequence Ratio (LCSR):** divide the length of the longest common subsequence by the length of the longest string.
- **Dice coefficient:**  $\frac{2 \times |n\text{-grams}(x) \cap n\text{-grams}(y)|}{|n\text{-grams}(x)| + |n\text{-grams}(y)|}$ .
- ...



## Generating/filtering the cognate list

- Train the C-SMT model on pairs identified through orthographic similarity. Generate new pairs for each word in the  $L1$  vocabulary.
- We then combine some heuristics to filter out bad pairs:
  - The C-SMT system gives a confidence score  $C$  to each translation.
  - Cognate pairs with very different frequencies are often wrong.
  - Cognate pairs should occur in similar contexts:

<b>4-gram</b>	$w_1$	$w_2$	$w_3$	$w_4$
		↕	↕	
	$v_1$	$v_2$	$v_3$	$v_4$
<b>Example</b>	diferència	de	càrrega	elèctrica
<b>CA-ES</b>		↕	↕	
	diferencia	de	carga	eléctrica

# Generation of the POS-annotated corpus

- Transfer most frequent POS for word  $w$  in  $L1$  to its translation in  $L2$ .
- For words left out in  $L2$ , use suffix analogy to known words to infer a POS.
- Accuracy up to 91.6%, but worse for Germanic languages.

	Total	# Tags
AN←ES	85.4%	42
CA←ES 500k	85.9%	42
CA←ES 140M	89.1%	42
NL←DE	59.0%	55
PFL←DE	65.1%	55
HSB←CS	83.6%	57
SK←CS	91.6%	57
PL←CS	77.6%	57