Quality of Parallel Crawled Data: Translationese, Machinese, Transcreations

SMART-Select Workshop on Data Curation for (Neural) Machine Translation

20/11/2018

Antonio Toral

university of
groningen

Contents

1. Quantity... or back in the SMT era
2. Quality and Translationese
3. Ebooks and Transcreations
4. Ideas and discussion

Acks

- Massive Acquisition
- Cleaning corpora

- Filip Klubicka
- Gema Ramirez-Sanchez
- Mikel Forcada
- Miquel Esplà
- Nikola Ljubesic
- Prokopis Prokopidis
- Raphael Rubino
- Sergio Ortiz-Rojas
- Tommi Pirinen
- Vassilis Papavassiliou
- Víctor M. Sánchez-Cartagena

MARIE CURIE ACTIONS

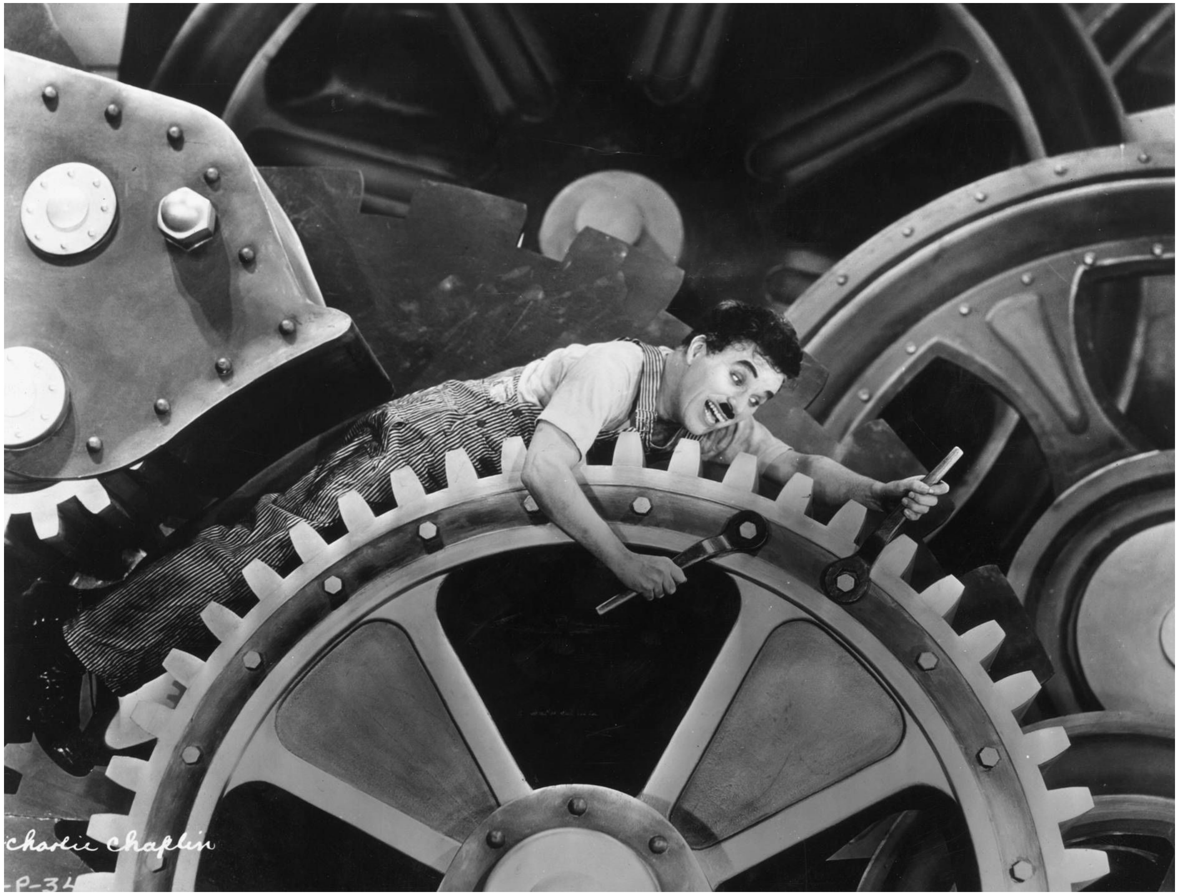
АБУМАТРАН
ABUMATRAN

- Quality
- Transcreations

- Andy Way
- Ian Matroos
- Joss Moorkens
- Ke Hu
- Sheila Castilho

PiPeNovel

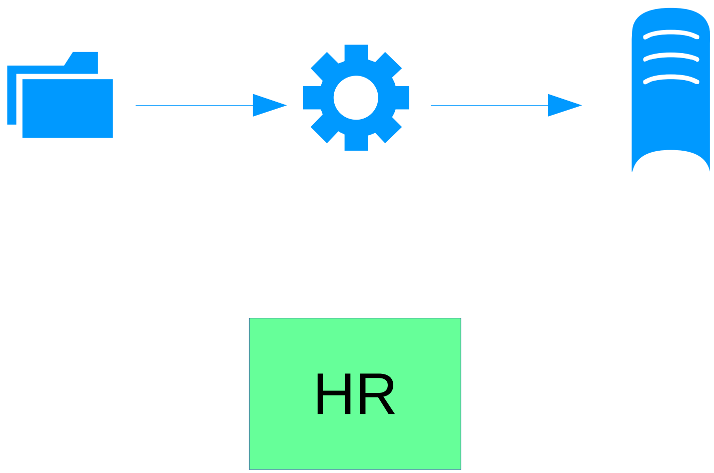
EUROPEAN ASSOCIATION FOR MACHINE TRANSLATION



1 Quantity… SMT era

Automatic Building of MT (2013-16)

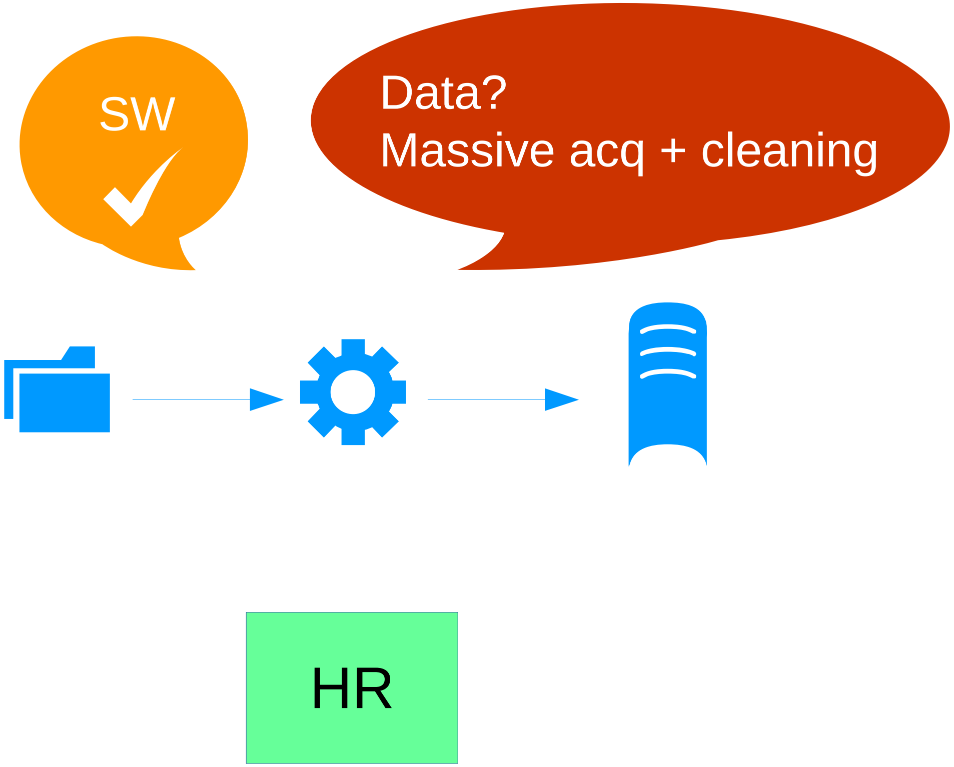
АБУМАТРАН
ABUMATRAN



HR

Automatic Building of MT (2013-16)

АБУМАТРАН
ABUMATRAN



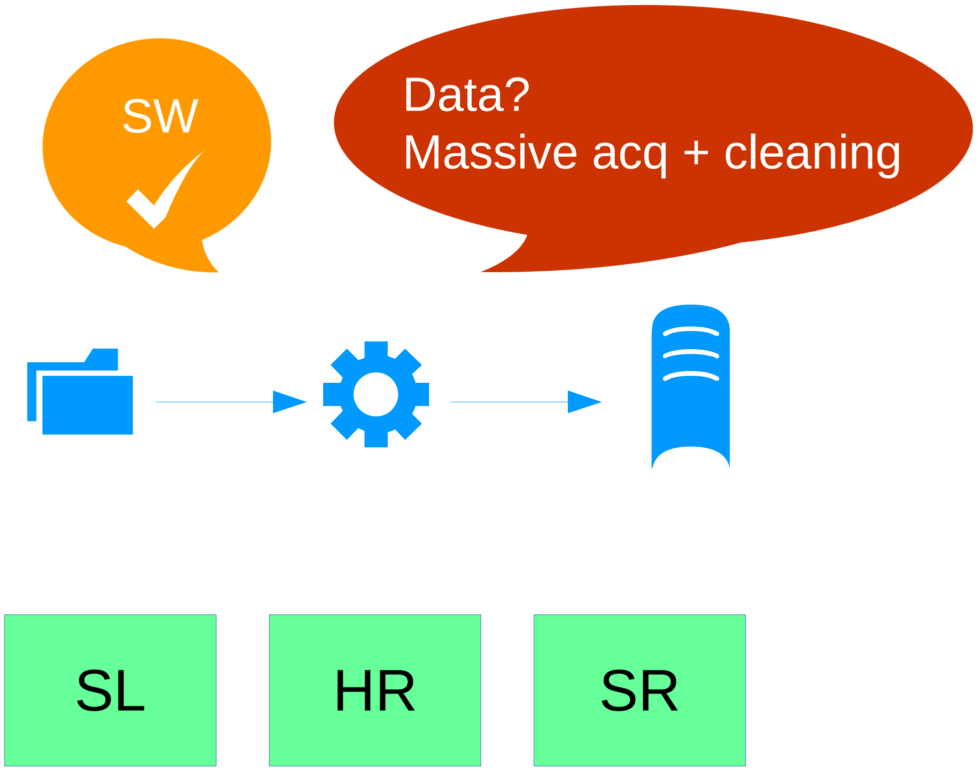
SW ✓

Data?
Massive acq + cleaning

HR

Automatic Building of MT (2013-16)

АБУМАТРАН
ABUMATRAN



SW ✓

Data?
Massive acq + cleaning

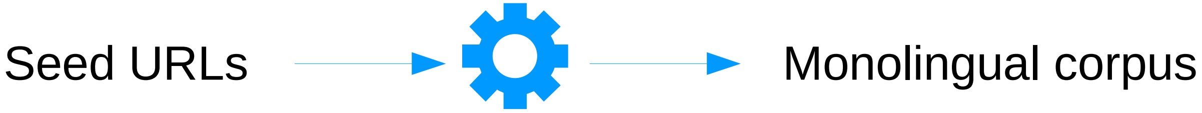
SL

HR

SR

Monolingual data (web)

- Motivation: Big LMs in SMT (Heafield et al., ACL’13)
- Massive crawling from TLDs with Spiderling



Seed URLs

Monolingual corpus

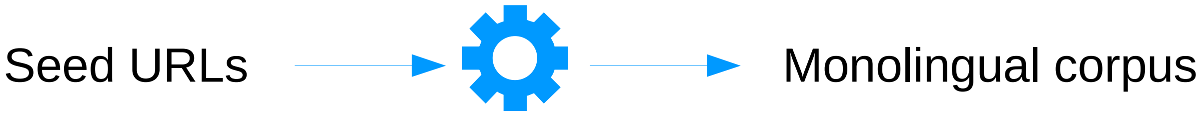
•

–

•

Monolingual data (web)

- Motivation: Big LMs in SMT (Heafield et al., ACL’13)
- Massive crawling from TLDs with Spiderling



Seed URLs

Monolingual corpus

- ~ 2 weeks → 1 billion words
  - HrWaC (Ljubešić & Erjavec, TSD’11), caWaC (Ljubešić & Toral, LREC’14), etc.
- Still useful for NMT?

Monolingual data (Twitter)

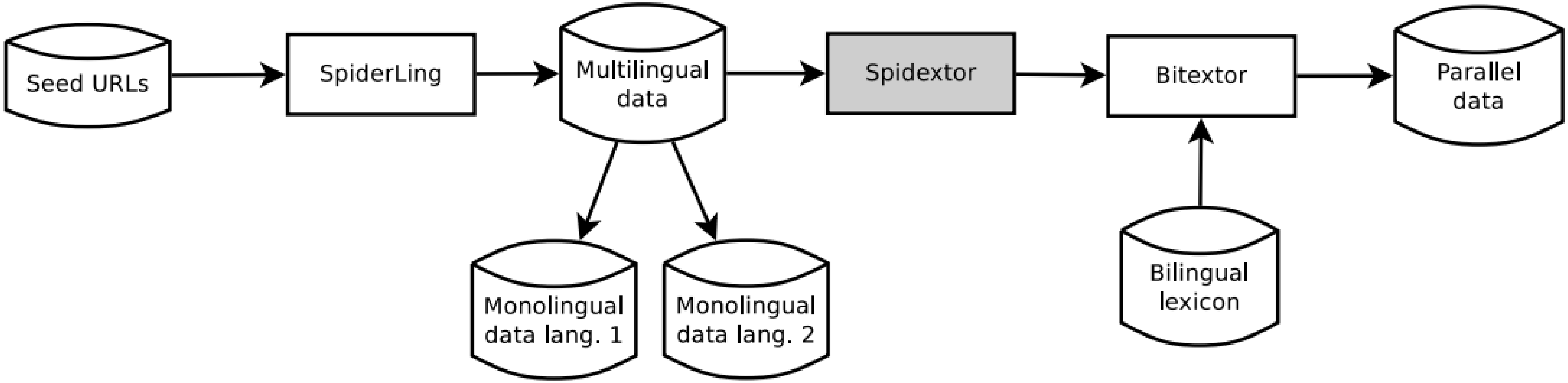
- Motivation
  - Cheap domain adaptation
  - Scarcity of parallel data

- Tool: TweetCat (Ljubešić et al., 2014)
  - Crawl tweets, tailored for *small* languages

- Application: Tweet MT (Toral et al., 2015)
  - CA, ES, EU, GL, PT

Parallel Data

- Spidextor: joint crawl of mono and parallel data from TLDs (Ljubešić et al, LREC’16)



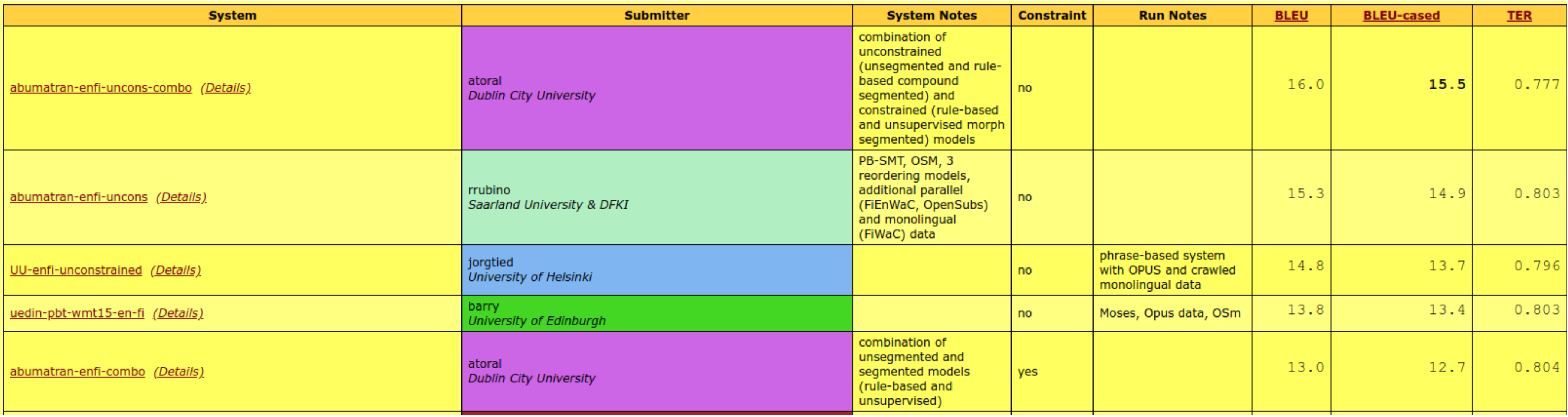
Parallel Data

- Spidextor: joint crawl of mono and parallel data from TLDs (Ljubešić et al, LREC’16)

Language Pair	**Crawling time**	**# segments**	**# words**
EN--FI	7 days	4M	100M
EN--HR	NA	2.4M	72M
EN--SL	3 days	1M	38M
EN--SR	NA	0.6M	27M

Parallel Data

- Use in MT (Rubino et al., WMT'15)
  - Crawling
    - Monolingual: Spiderling
    - Parallel: Bitextor + ILSP-FC



System	Submitter	System Notes	Constraint	Run Notes	BLEU	BLEU-cased	TER
abumatran-enfi-uncons-combo *(Details)*	atoral *Dublin City University*	combination of unconstrained (unsegmented and rule-based compound segmented) and constrained (rule-based and unsupervised morph segmented) models	no		16.0	**15.5**	0.777
abumatran-enfi-uncons *(Details)*	rrubino *Saarland University & DFKI*	PB-SMT, OSM, 3 reordering models, additional parallel (FiEnWaC, OpenSubs) and monolingual (FiWaC) data	no		15.3	14.9	0.803
UU-enfi-unconstrained *(Details)*	jorgtied *University of Helsinki*		no	phrase-based system with OPUS and crawled monolingual data	14.8	13.7	0.796
uedin-pbt-wmt15-en-fi *(Details)*	barry *University of Edinburgh*		no	Moses, Opus data, OSm	13.8	13.4	0.803
abumatran-enfi-combo *(Details)*	atoral *Dublin City University*	combination of unsegmented and segmented models (rule-based and unsupervised)	yes		13.0	12.7	0.804

Source: http://matrix.statmt.org/matrix/systems_list/1775

Cleaning Noisy Corpora

- Many publicly available parallel corpora are potentialy useful
- But... they are too noisy
  - Missalignments
  - Encoding errors
  - etc
- E.g. OpenSubtitles

Cleaning Noisy Corpora

- Automatic cleaning (Forcada et al., 2014)
  - Fixing (sparsity)
  - Removing sentences (noise)

Cleaning Noisy Corpora

- Automatic cleaning (Forcada et al., 2014)
  - Fixing (sparsity)
    - Converting Cyrillic characters to their Latin counterparts
    - Converting encoding to UTF-8
    - Spelling errors
    - Inconsistent punctuation marks, numbers and spacing
  - Removing sentences (noise)
    - Without alphabetical characters
    - Too different in length
    - Not in the right language

Cleaning Noisy Corpora

- Data
  - Corpora: OpenSubtitles EN—HR
  - Input: 30M sentence pairs
  - Output: 17M

- Extrinsic Evaluation
  - Train MT system with OpenSubs as is vs cleaned
  - Test set: news domain (WMT13)

Cleaning Noisy Corpora

- SMT results (BLEU)

	EN-to-HR	HR-to-EN
OpenSubs as is	0.09	0.22
OpenSubs cleaned	0.22	0.31
Relative improvement	145%	37%

- Use for NMT: dedicated shared task at WMT18

2 Quality and Translationese

Quality and Translationese

- MT performs better if training data consists on original SL text translated directly into TL (Kurokawa et al., 2009)
  - But that is not how MT practitioners use corpora, e.g. Europarl

-
  -
  -

Quality and Translationese

- MT performs better if training data consists on original SL text translated directly into TL (Kurokawa et al., 2009)
  - But that is not how MT practitioners use corpora, e.g. Europarl

- Idea: given a crawled document, identify:
  - Original or translationese
  - If translationese, its original language

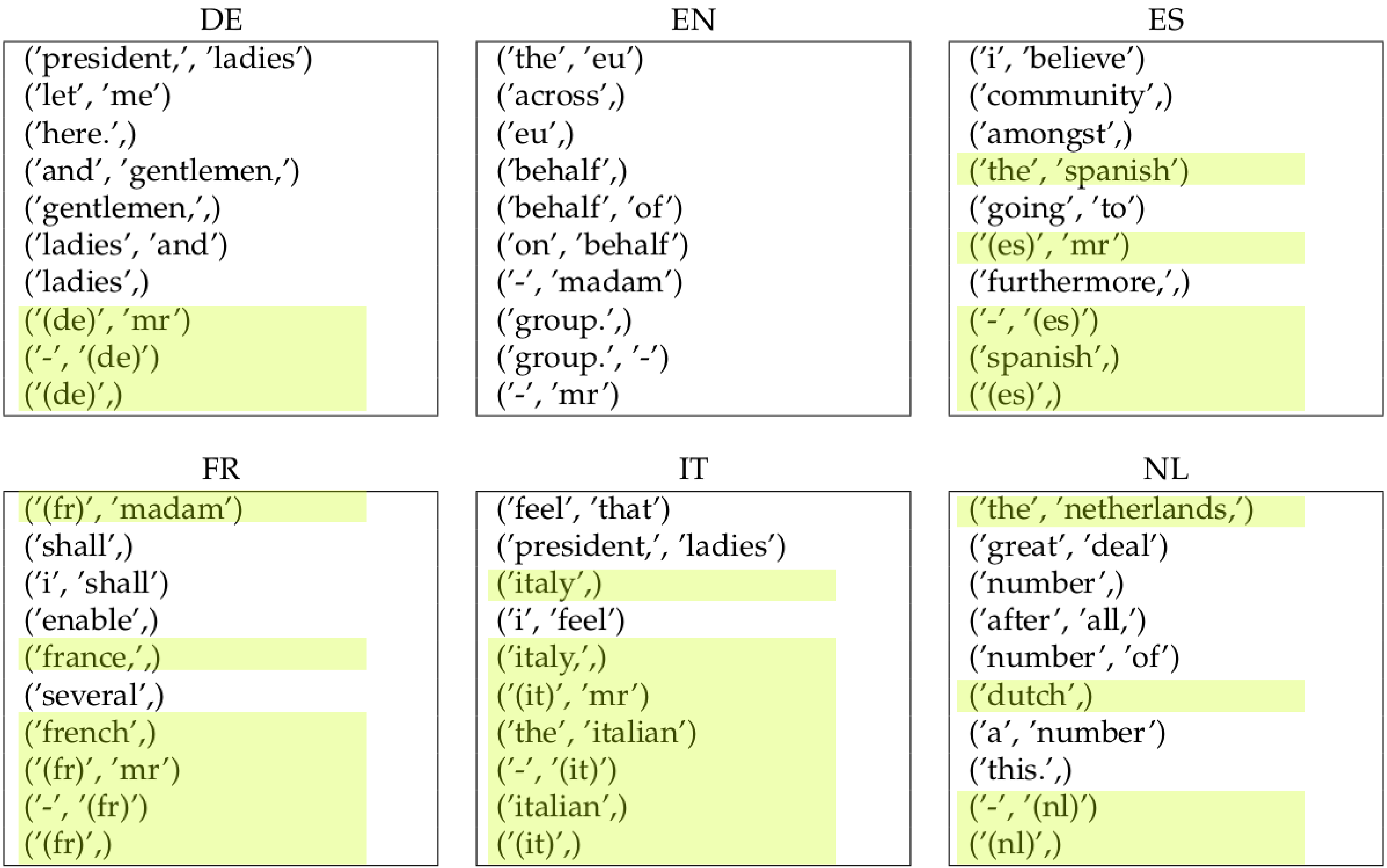
Source language identification

- Halteren (2008): token-based features
  - Up to 87% accuracy on Europarl

- Koppel and Ordan (2011): function words
  - 93% accuracy on Europarl, 65% out-domain (news)

- Matroos (2018): PoS tags
  - Works out-of-the-box for the 73 languages in UD
  - Vs Halteren (2008)
    - Worse on in-domain (Europarl) → 0.69 vs 0.88
    - Better on out-domain (Books) → 0.74 vs 0.69

Token-based features



DE

(’president,’, ’ladies’)
(’let’, ’me’)
(’here.’,)
(’and’, ’gentlemen,’)
(’gentlemen,’,)
(’ladies’, ’and’)
(’ladies’,)
(’(de)’, ’mr’)
(’-’, ’(de)’)
(’(de)’,)

EN

(’the’, ’eu’)
(’across’,)
(’eu’,)
(’behalf’,)
(’behalf’, ’of’)
(’on’, ’behalf’)
(’-’, ’madam’)
(’group.’,)
(’group.’, ’-’)
(’-’, ’mr’)

ES

(’i’, ’believe’)
(’community’,)
(’amongst’,)
(’the’, ’spanish’)
(’going’, ’to’)
(’(es)’, ’mr’)
(’furthermore,’,)
(’-’, ’(es)’)
(’spanish’,)
(’(es)’,)

FR

(’(fr)’, ’madam’)
(’shall’,)
(’i’, ’shall’)
(’enable’,)
(’france,’,)
(’several’,)
(’french’,)
(’(fr)’, ’mr’)
(’-’, ’(fr)’)
(’(fr)’,)

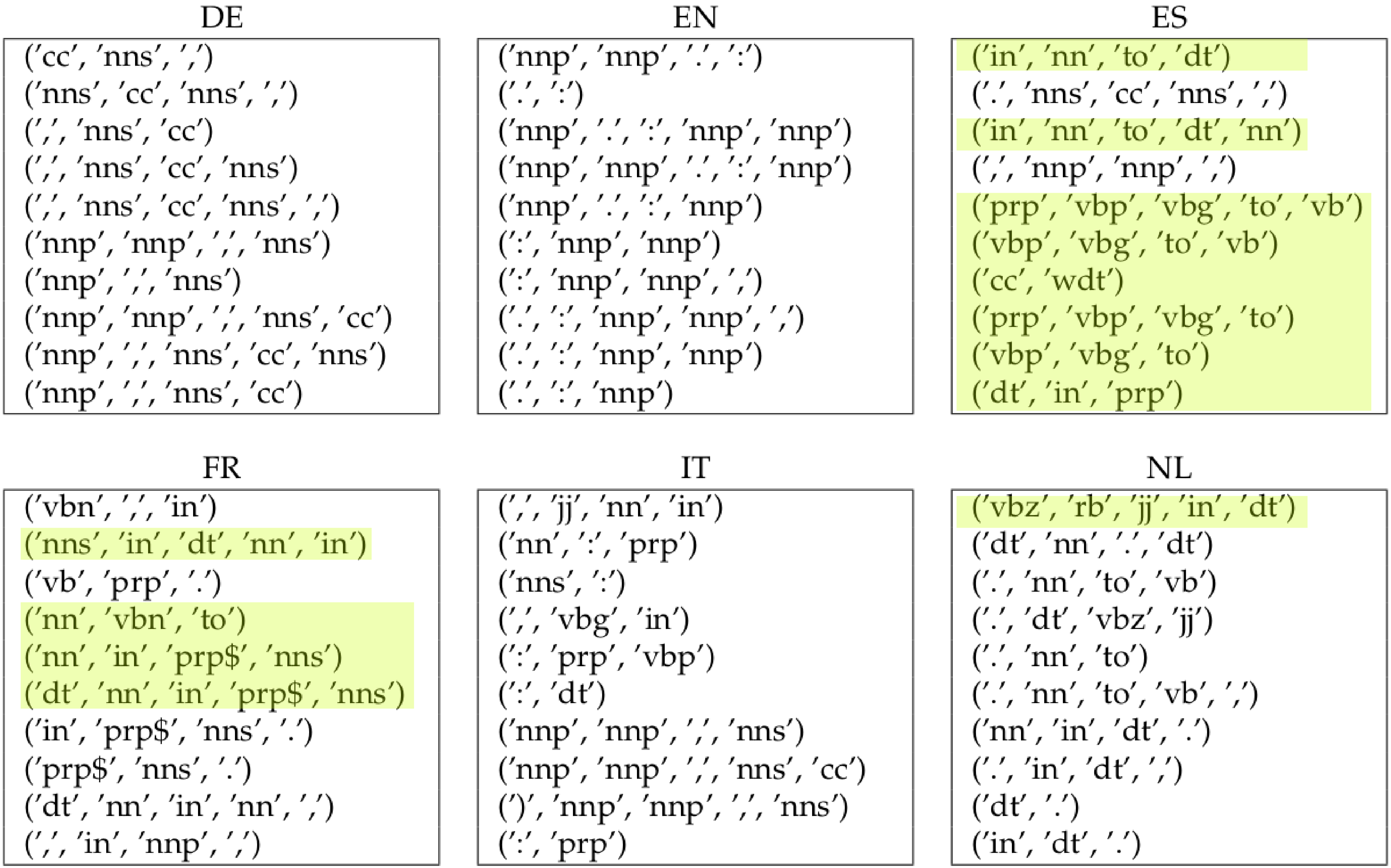
IT

(’feel’, ’that’)
(’president,’, ’ladies’)
(’italy’,)
(’i’, ’feel’)
(’italy,’,)
(’(it)’, ’mr’)
(’the’, ’italian’)
(’-’, ’(it)’)
(’italian’,)
(’(it)’,)

NL

(’the’, ’netherlands,’)
(’great’, ’deal’)
(’number’,)
(’after’, ’all,’)
(’number’, ’of’)
(’dutch’,)
(’a’, ’number’)
(’this.’,)
(’-’, ’(nl)’)
(’(nl)’,)

PoS-based features

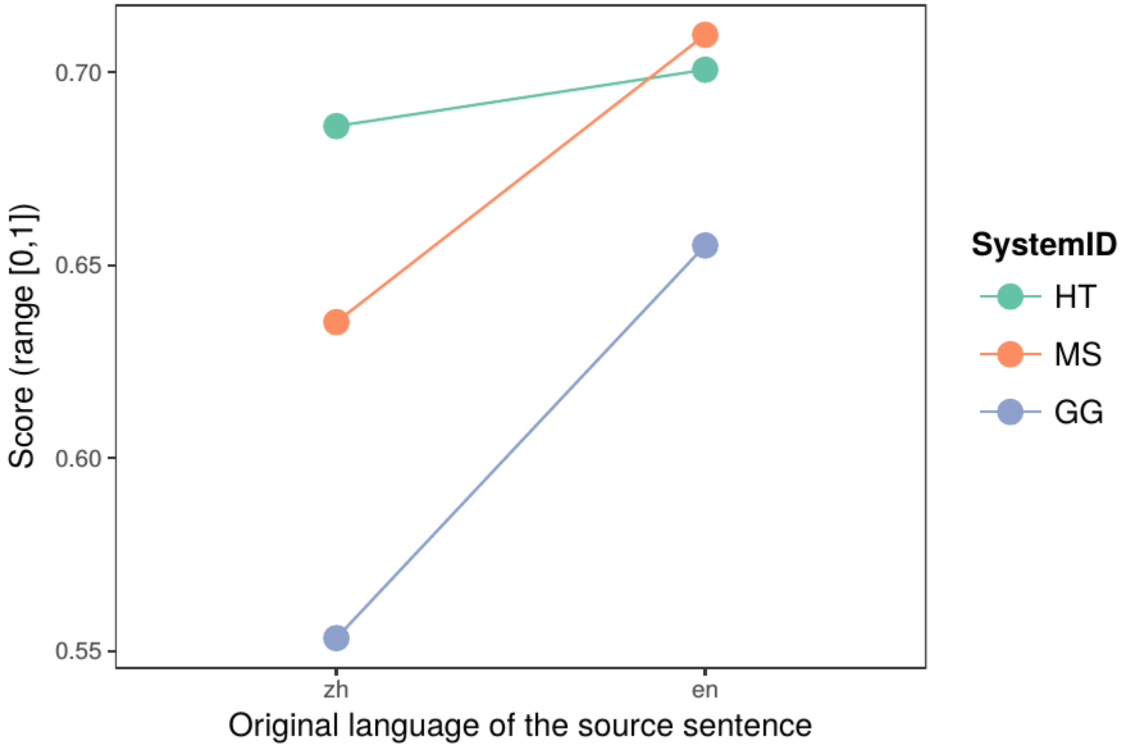


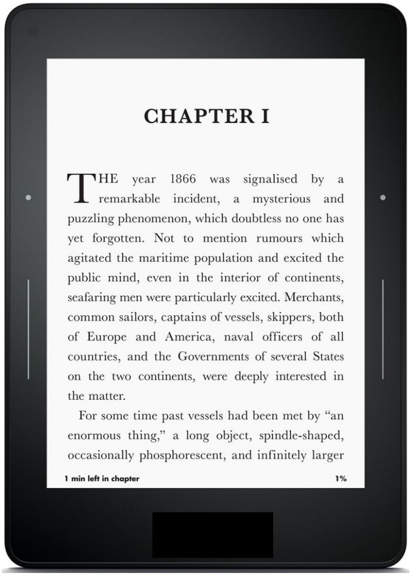
DE	EN	ES
(’cc’, ’nns’, ’,’)	(’nnp’, ’nnp’, ’.’, ’:’)	(’in’, ’nn’, ’to’, ’dt’)
(’nns’, ’cc’, ’nns’, ’,’)	(’.’, ’:’)	(’.’, ’nns’, ’cc’, ’nns’, ’,’)
(’,’, ’nns’, ’cc’)	(’nnp’, ’.’, ’:’, ’nnp’, ’nnp’)	(’in’, ’nn’, ’to’, ’dt’, ’nn’)
(’,’, ’nns’, ’cc’, ’nns’)	(’nnp’, ’nnp’, ’.’, ’:’, ’nnp’)	(’,’, ’nnp’, ’nnp’, ’,’)
(’,’, ’nns’, ’cc’, ’nns’, ’,’)	(’nnp’, ’.’, ’:’, ’nnp’)	(’prp’, ’vbp’, ’vbg’, ’to’, ’vb’)
(’nnp’, ’nnp’, ’,’, ’nns’)	(’:’, ’nnp’, ’nnp’)	(’vbp’, ’vbg’, ’to’, ’vb’)
(’nnp’, ’,’, ’nns’)	(’:’, ’nnp’, ’nnp’, ’,’)	(’cc’, ’wdt’)
(’nnp’, ’nnp’, ’,’, ’nns’, ’cc’)	(’.’, ’:’, ’nnp’, ’nnp’, ’,’)	(’prp’, ’vbp’, ’vbg’, ’to’)
(’nnp’, ’,’, ’nns’, ’cc’, ’nns’)	(’.’, ’:’, ’nnp’, ’nnp’)	(’vbp’, ’vbg’, ’to’)
(’nnp’, ’,’, ’nns’, ’cc’)	(’.’, ’:’, ’nnp’)	(’dt’, ’in’, ’prp’)

FR	IT	NL
(’vbn’, ’,’, ’in’)	(’,’, ’jj’, ’nn’, ’in’)	(’vbz’, ’rb’, ’jj’, ’in’, ’dt’)
(’nns’, ’in’, ’dt’, ’nn’, ’in’)	(’nn’, ’:’, ’prp’)	(’dt’, ’nn’, ’.’, ’dt’)
(’vb’, ’prp’, ’.’)	(’nns’, ’:’)	(’.’, ’nn’, ’to’, ’vb’)
(’nn’, ’vbn’, ’to’)	(’,’, ’vbg’, ’in’)	(’.’, ’dt’, ’vbz’, ’jj’)
(’nn’, ’in’, ’prp$’, ’nns’)	(’:’, ’prp’, ’vbp’)	(’.’, ’nn’, ’to’)
(’dt’, ’nn’, ’in’, ’prp$’, ’nns’)	(’:’, ’dt’)	(’.’, ’nn’, ’to’, ’vb’, ’,’)
(’in’, ’prp$’, ’nns’, ’.’)	(’nnp’, ’nnp’, ’,’, ’nns’)	(’nn’, ’in’, ’dt’, ’.’)
(’prp$’, ’nns’, ’.’)	(’nnp’, ’nnp’, ’,’, ’nns’, ’cc’)	(’.’, ’in’, ’dt’, ’,’)
(’dt’, ’nn’, ’in’, ’nn’, ’,’)	(’)’, ’nnp’, ’nnp’, ’,’, ’nns’)	(’dt’, ’.’)
(’,’, ’in’, ’nnp’, ’,’)	(’:’, ’prp’)	(’in’, ’dt’, ’.’)

Translationese in Test

- Reassessing human parity (Toral et al., WMT'18)



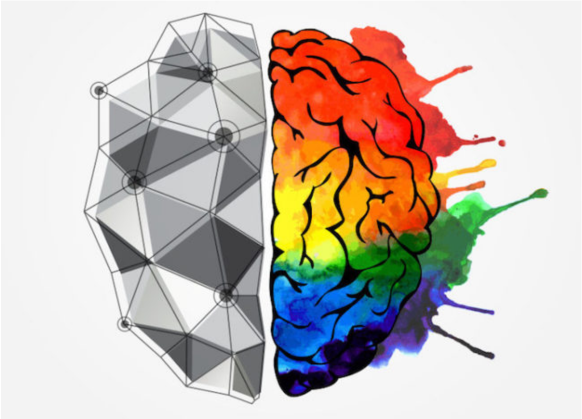


CHAPTER I

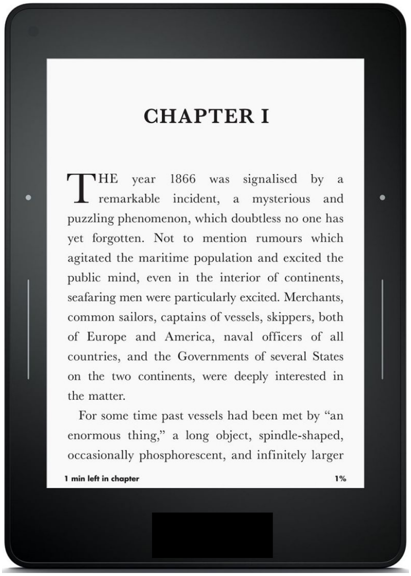
THE year 1866 was signalised by a remarkable incident, a mysterious and puzzling phenomenon, which doubtless no one has yet forgotten. Not to mention rumours which agitated the maritime population and excited the public mind, even in the interior of continents, seafaring men were particularly excited. Merchants, common sailors, captains of vessels, skippers, both of Europe and America, naval officers of all countries, and the Governments of several States on the two continents, were deeply interested in the matter.

For some time past vessels had been met by "an enormous thing," a long object, spindle-shaped, occasionally phosphorescent, and infinitely larger

1 min left in chapter 1%



3 Ebooks and Transcreations



CHAPTER I

THE year 1866 was signalised by a remarkable incident, a mysterious and puzzling phenomenon, which doubtless no one has yet forgotten. Not to mention rumours which agitated the maritime population and excited the public mind, even in the interior of continents, seafaring men were particularly excited. Merchants, common sailors, captains of vessels, skippers, both of Europe and America, naval officers of all countries, and the Governments of several States on the two continents, were deeply interested in the matter.

For some time past vessels had been met by "an enormous thing," a long object, spindle-shaped, occasionally phosphorescent, and infinitely larger

1 min left in chapter 1%

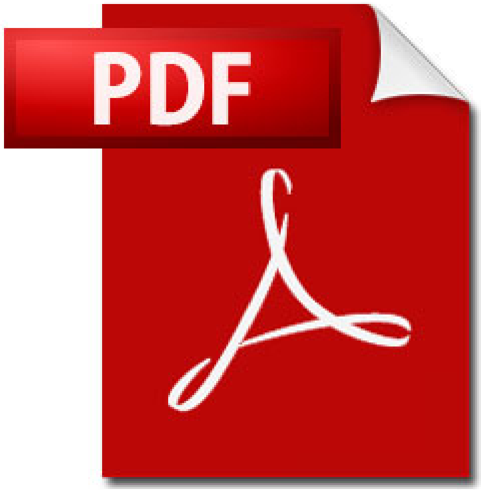
[redacted]

Ebooks as a source to crawl parallel data?

Question

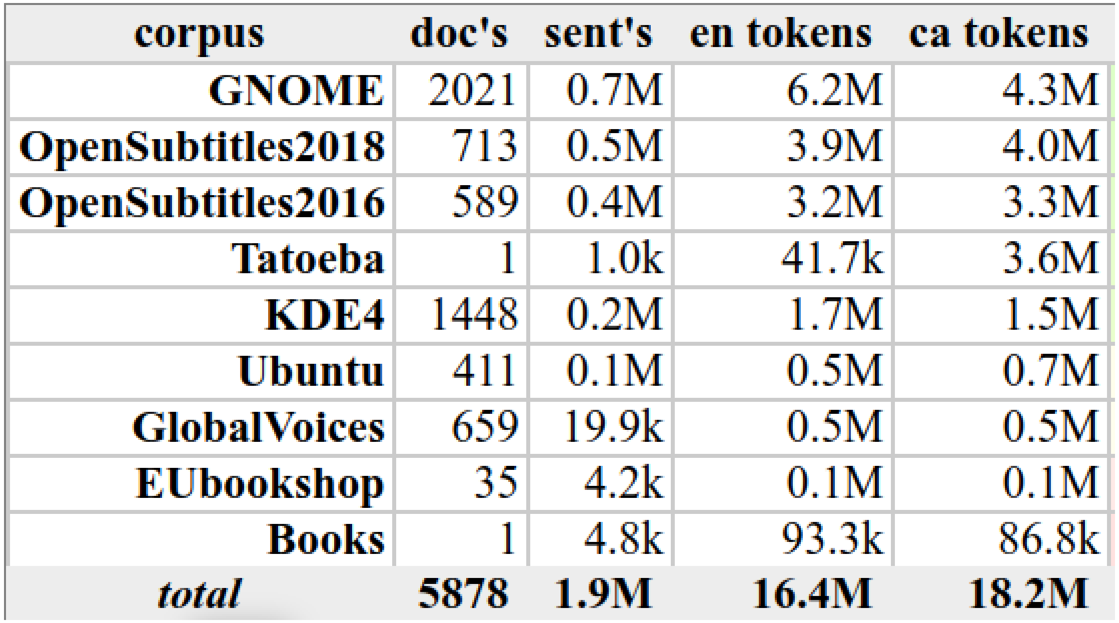
Parallel Data from Ebooks

≠



Motivation

- Literary-adapted MT for EN→CA (Toral and Way, 2018)



corpus	doc's	sent's	en tokens	ca tokens
GNOME	2021	0.7M	6.2M	4.3M
OpenSubtitles2018	713	0.5M	3.9M	4.0M
OpenSubtitles2016	589	0.4M	3.2M	3.3M
Tatoeba	1	1.0k	41.7k	3.6M
KDE4	1448	0.2M	1.7M	1.5M
Ubuntu	411	0.1M	0.5M	0.7M
GlobalVoices	659	19.9k	0.5M	0.5M
EUbookshop	35	4.2k	0.1M	0.1M
Books	1	4.8k	93.3k	86.8k
total	**5878**	**1.9M**	**16.4M**	**18.2M**

EN—CA corpora on http://opus.nlpl.eu/

Pipeline

Given an ebook in EN and its translation in CA

1. Epub (or mobi) to text	Calibre tools
2. Normalisation	Moses
3. Sentence splitting	NLTK/Freeling
4. Sentence alignment	Hunalign, Apertium dict

Result

- Training
  - Parallel: 133 book pairs
    - 1.2M sentence pairs
  - Mono: 1,000 books
    - >5M sentences
- Test
  - 12 books: 86K sentence pairs

Result

- Advantages
  - Clean data and easy to process. EPUB ≠ PDF
  - High quality translations
  - Present day language (vs Gutenberg)

- Disadvantages
  - Tedious: find and buy books, DRM, …
  - Copyright

Open Questions

- Can this be useful...
  - … as out-domain data? How domain-specific is it?
  - … for better resourced language pairs?

Translation Options

French

J’étais épuisé et je me suis jeté sur ma couchette.
Je crois que j’ai dormi parce que je me suis réveillé avec des étoiles sur le visage.

English – Prof. Translation 1

But all this excitement had exhausted me and I dropped heavily on to my sleeping plank.
I must have had a longish sleep, for, when I woke, the stars were shining down on my face.

English – Prof. Translation 2

I was exhausted and threw myself on my bunk.
I must have fallen asleep, because I woke up with the stars in my face.

Which translation do you prefer?

Translation Options

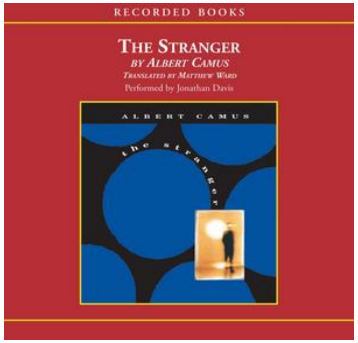
French J’étais épuisé et je me suis jeté sur ma couchette. Je crois que j’ai dormi parce que je me suis réveillé avec des étoiles sur le visage.
English – Gilbert (1946) But all this excitement had exhausted me and I dropped heavily on to my sleeping plank. I must have had a longish sleep, for, when I woke, the stars were shining down on my face.
English – Ward (1989) I was exhausted and threw myself on my bunk. I must have fallen asleep, because I woke up with the stars in my face.

ALBERT Camus

THE Stranger

TRANSLATED FROM
THE FRENCH
BY STUART GILBERT

NEW YORK
VINTAGE BOOKS
1958



Translation Options

French

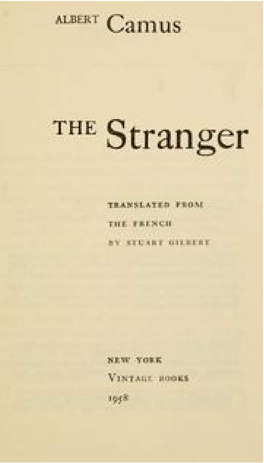
J’étais épuisé et je me suis jeté sur ma couchette.
Je crois que j’ai dormi parce que je me suis réveillé avec des étoiles sur le visage.

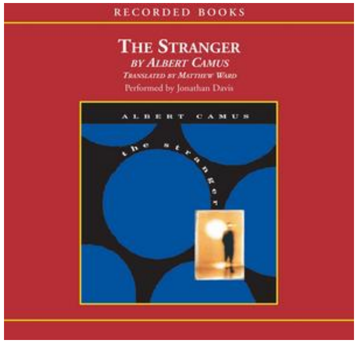
English – Gilbert (1946)

But all this excitement had exhausted me and I dropped heavily on to my sleeping plank.
I must have had a longish sleep, for, when I woke, the stars were shining down on my face.

English – Ward (1989)

I was exhausted and threw myself on my bunk.
I must have fallen asleep, because I woke up with the stars in my face.





Translation Options

French

J’étais épuisé et je me suis jeté sur ma couchette.
Je crois que j’ai dormi parce que je me suis réveillé avec des étoiles sur le visage.

English – Gilbert (1946)

But all this excitement had exhausted me and I dropped heavily on to my sleeping plank.
I must have had a longish sleep, for, when I woke, the stars were shining down on my face.

Domesticating
Transcreation
Free translation

English – Ward (1989)

I was exhausted and threw myself on my bunk.
I must have fallen asleep, because I woke up with the stars in my face.

Foreignising
Literal translation

Translation Options

French

J’étais épuisé et je me suis jeté sur ma couchette.
Je crois que j’ai dormi parce que je me suis réveillé avec des étoiles sur le visage.

English – Gilbert (1946)

But all this excitement had exhausted me and I dropped heavily on to my sleeping plank.
I must have had a longish sleep, for, when I woke, the stars were shining down on my face.

BLEU 0.11
TER 0.80

English – Ward (1989)

I was exhausted and threw myself on my bunk.
I must have fallen asleep, because I woke up with the stars in my face.

BLEU 0.28
TER 0.56

Translation Options

- A human translation falls somewhere between
  - Domesticated / transcreation / free translation
  - Foreignising / literal
- Which school of thought is prevalent nowadays?
- Is this important when crawling data?

4 Ideas and Discussion

Ideas and Discussion

- Monolingual data
  - Not (that) important anymore with NMT?
    - Bracktranslate vs unsupervised NMT

- Quality
  - Filtering (dedicated shared task at WMT’18)
  - Translation options
  - Identification
    - Original Language
    - Translated? Human- or machine-translated?
      - Classifiers worked well to identify translations by SMT, but NMT output is more fluent and impredictable...

Quantity or Quality?

~~Quantity or Quality?~~

Quantity **and** Quality

Quantity and quality

- Quantity: crawl as much as possible
- Quality
  - Filter out
    - Not parallel, dirty, etc
    - MT
  - Augment crawled data with metadata
    - Translationese: original or translated (+ confidence)
    - If translated → original language (+ confidence)
    - Translation type → from literal to transcreation (continuous)
    - Provenance → domain information (Tars and Fishel, 2018)

Thanks!
Questions?

Antonio Toral
a.toral.ruiz@rug.nl
@_atoral

References

- M. Forcada, S. Ortiz-Rojas, T. Pirinen, R. Rubino, A. Toral. 2014. Abu-MaTran Deliverable D4.1b MT systems for the second development cycle.
- H. Halteren. 2008. Source language markers in europarl translations. COLING.
- K. Heafield, I. Pouzyrevsky, J. H. Clark, P. Koehn. 2013. Scalable Modified Kneser-Ney Language Model Estimation. ACL
- M. Koppel, N. Ordan. 2011. Translationese and its dialects. ACL.
- D. Kurokawa, C. Goutte, P. Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. MT-Summit.
- N. Ljubešić, T. Erjavec. 2011. hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. TSD.
- N. Ljubešić, M. Esplà-Gomis, A. Toral, S. Ortiz-Rojas, F. Klubička. 2016. Producing Monolingual and Parallel Web Corpora at the Same Time – SpiderLing and Bitextor’s Love Affair. LREC.
- N. Ljubešić, D. Fišer, T. Erjavec. 2014. TweetCaT: a tool for building Twitter corpora of smaller languages. LREC.

References

- N. Ljubešić, A. Toral. 2014. caWaC – a Web Corpus of Catalan and its Application to Language Modeling and Machine Translation. LREC.
- I. Matroos. 2018. Source language prediction: A part of speech based approach. Bachelor Thesis. https://github.com/imatr/Source-language-prediction
- R. Rubino, T. Pirinen, M. Esplà-Gomis, N. Ljubešić, S. Ortiz-Rojas, V. Papavassiliou, P. Prokopidis, A. Toral. 2015. Abu-MaTran at WMT 2015 Translation Task: Morphological Segmentation and Web Crawling.
- S. Tars, M. Fishel. 2018. Multi-Domain Neural Machine Translation. EAMT.
- A. Toral, X. Wu, T. Pirinen, Z. Qiu, E. Bicici, J. Du. Dublin City University at the TweetMT 2015 Shared Task. SEPLN.
- A. Toral, S. Castilho, K. Hu, A. Way. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. WMT.
- A. Toral, A. Way. 2018. What level of quality can neural machine translation attain on literary text? Translation Quality Assessment: From Principles to Practice, Springer.