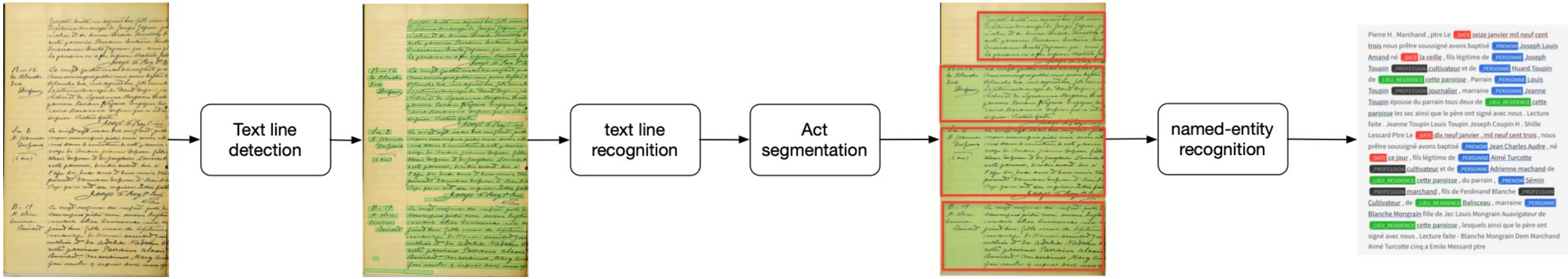
A Comprehensive Study of Open-source Libraries for Named Entity Recognition on Handwritten Historical Documents

Claire Bizon Monroc, Blanche Miret, Marie-Laurence Bonhomme, Christopher Kermorvant

TEKLIA

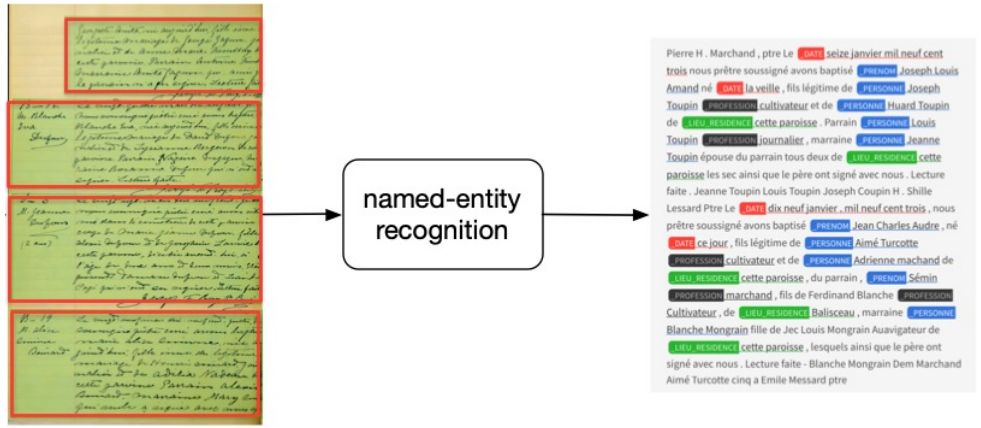
DAS 2022 – La Rochelle

Standard document processing workflow



If the final performance is not satisfactory,
on which step should we invest our effort/budget?

Improving NER



Benchmarks

Add a Result

These leaderboards are used to track progress in Named Entity Recognition

Trend	Dataset	Best Model	Paper	Code	Compare
	CoNLL 2003 (English)	ACE + document-context			See all
	Ontonotes v5 (English)	BERT-MRC+DSC			See all
	NCBI-disease	BioBERT			See all
	ACE 2005	Ours: cross-sentence ALB			See all
	WNUT 2017	CL-KL			See all
	SLUE	W2V2-L-LL60K (pipeline approach, uses LM)			See all
	BC5CDR	CL-L2			See all
	JNLPBA	KeBioLM			See all
	GENIA	Biaffine-NER			See all
	ACE 2004	Ours: cross-sentence ALB			See all

If NER has to be improved, should we invest on custom models or explore other libraries?

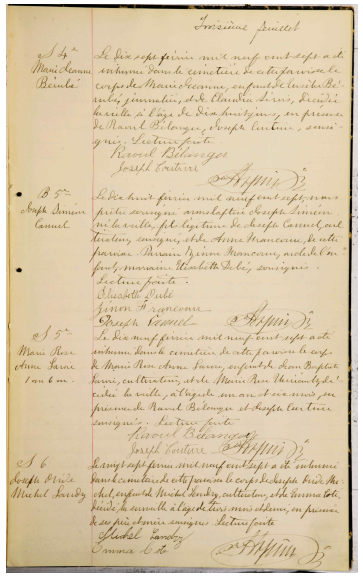
NER on historical documents

HOME

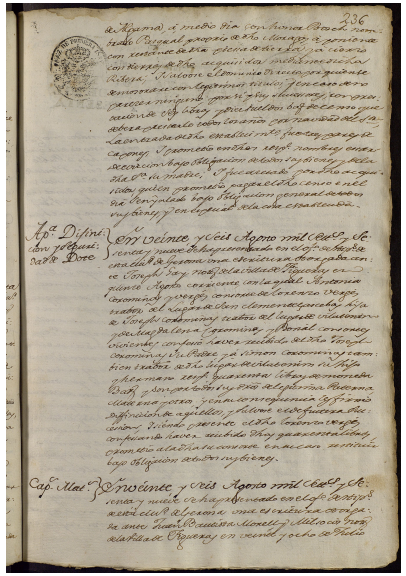
Medieval charters

Ancient form of language, (very) little punctuation and capitalisation

Balsac



Esposalles



Modern registers

Text is mostly entities, paragraphs with similar structures

Is the NER task different on historical documents compared to electronic documents ?

Open source NER libraries

spaCy

First “industrial” library for NER
Ecosystem of tools
Fast
Open source but a bit obscure

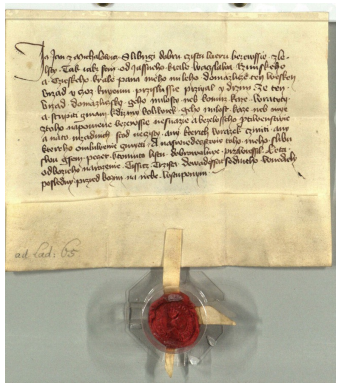
flair

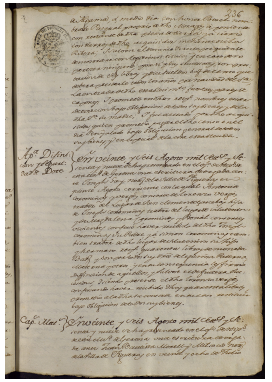
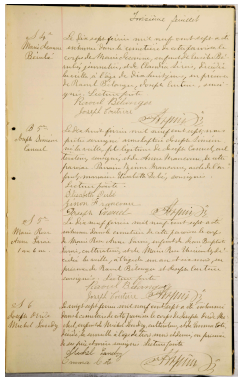
First to provide ready-to-use embedding
Simple
Based on pytorch
Models hosted on Huggingface 

Stanza

Evolution of Stanford CoreNLP with embeddings
66 languages (Latin, old French…)

Datasets



	HOME				Esposalles	Balsac	CoNLL
No. of	Czech	German	Latin	All	Catalan	French	Eng/Ger
Page	202	173	126	501	125	896	1,390
Line	3,591	3,199	1,971	8,761	3,827	45,479	
Word	66,257	77,086	35,759	179,102	39,527	205,165	301,418
Entity	4,117	4,419	3,315	11,851	16,782	25,564	35,089

Nested entities

Nos Fridericus, Dei gracia dux Austrie et Styrie [. . .] profitemur et recognoscimus
PER: Fridericus, Dei gracia dux Austrie et Styrie
LOC

- **Flatten approach**

  Nos Fridericus, Dei gracia dux Austrie et Styrie [. . .] profitemur et recognoscimus
  PER: Fridericus, Dei gracia dux
  LOC: Austrie - Styrie

- Hierarchical NER models

- Independent approach

  Nos Fridericus, Dei gracia dux Austrie et Styrie [. . .] profitemur et recognoscimus
  PER: Fridericus, Dei gracia dux Austrie et Styrie
  LOC: Austrie - Styrie

Metrics

Page level metrics

If an entity from the groundtruth is predicted with the true label and text, then it is considered as a true positive

Nerval for NER on automatic transcription

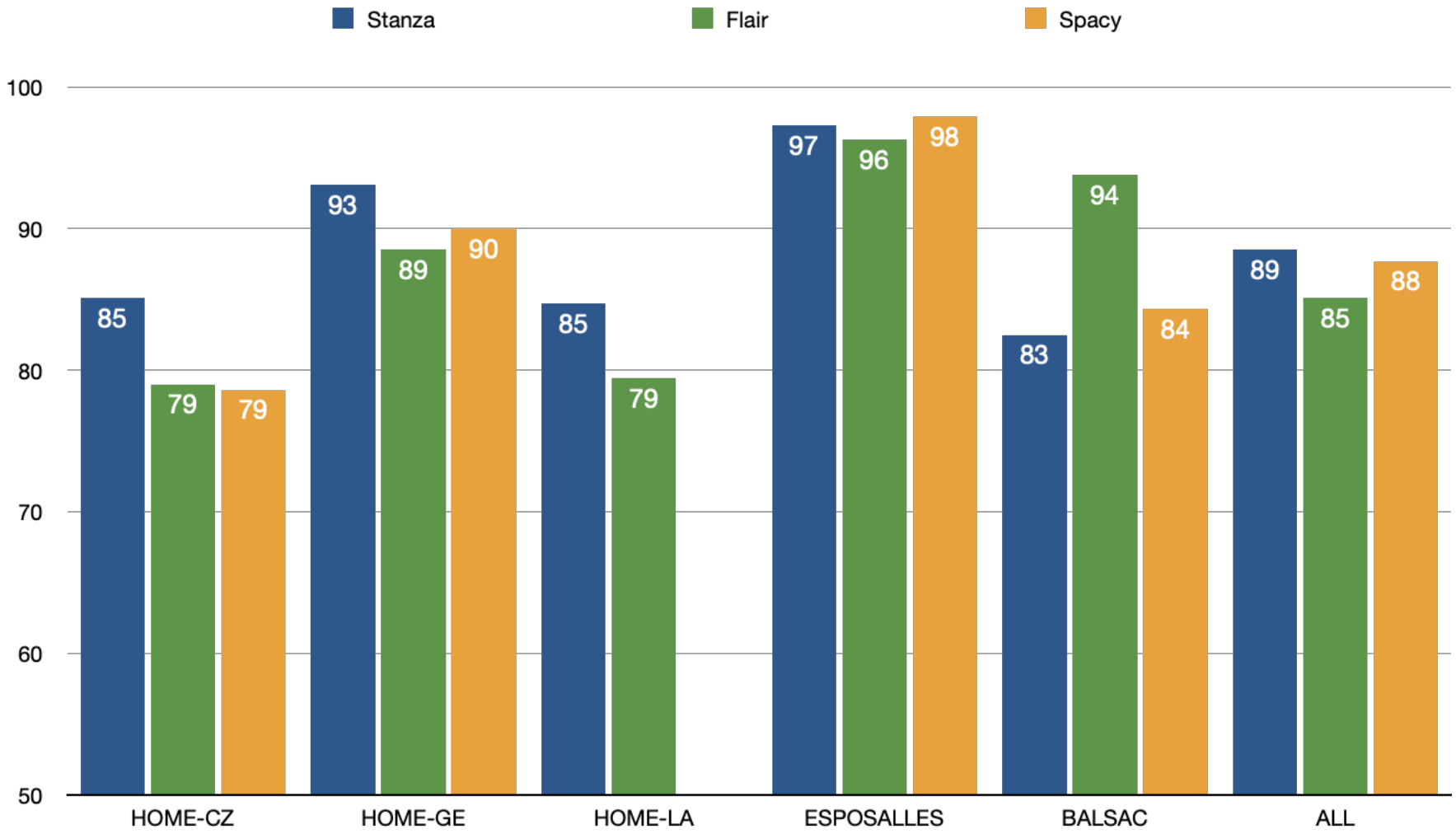
Open source python package for NER on noisy text https://gitlab.com/teklia/nerval

1. Aligment with groundtruth at character level
2. Aligment of entities (with text distance)
3. Compute Precision, Recall, F-score on Label

Esposalles metrics

1. Alignment of entities to groundtruth
2. Score the entities
   - Label mismatch = 0
   - Label match = 1 – CER
3. Compute accuracy

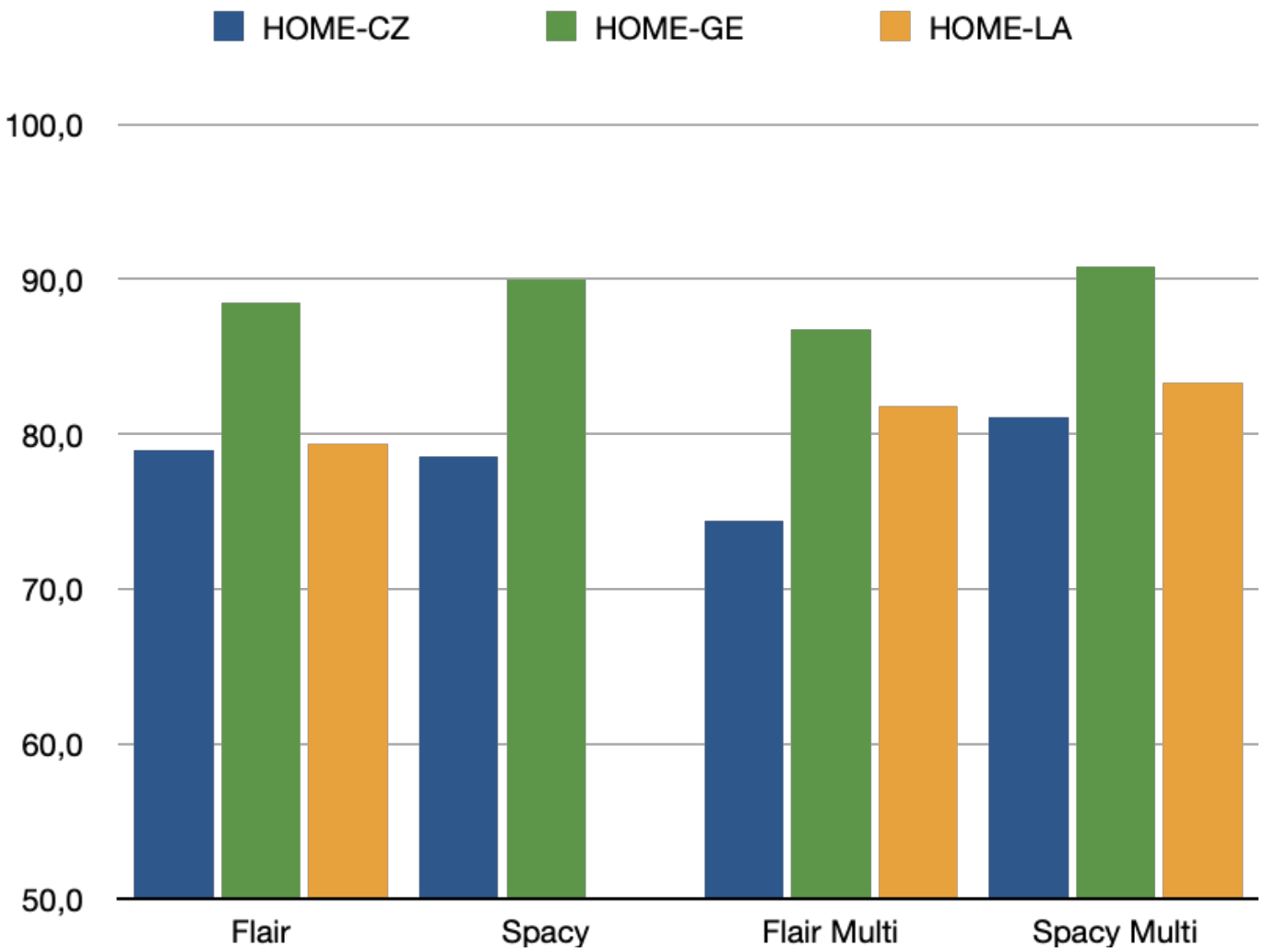
Results on manual transcriptions



No clear winner

Effect of careful hype-parameters optimization ?

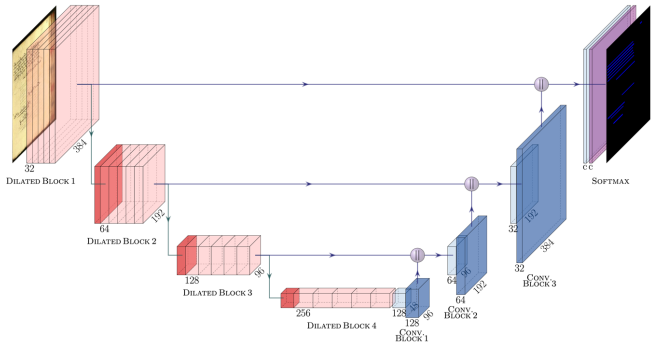
Results with multi-lingual models



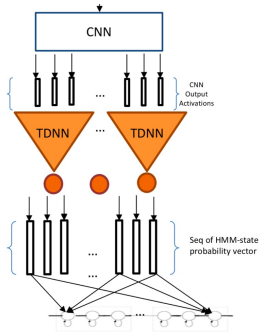
Spacy multi-lingual is better

More data is better than specialized models

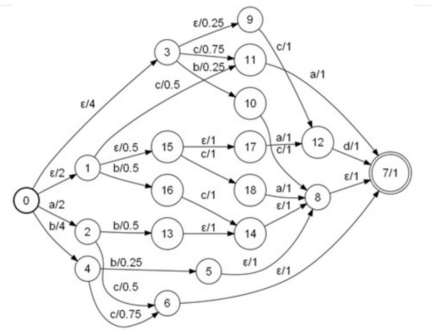
Handwritten Text Recognition



Text line detection with Doc-UFCN



Optical model CNN/TDNN

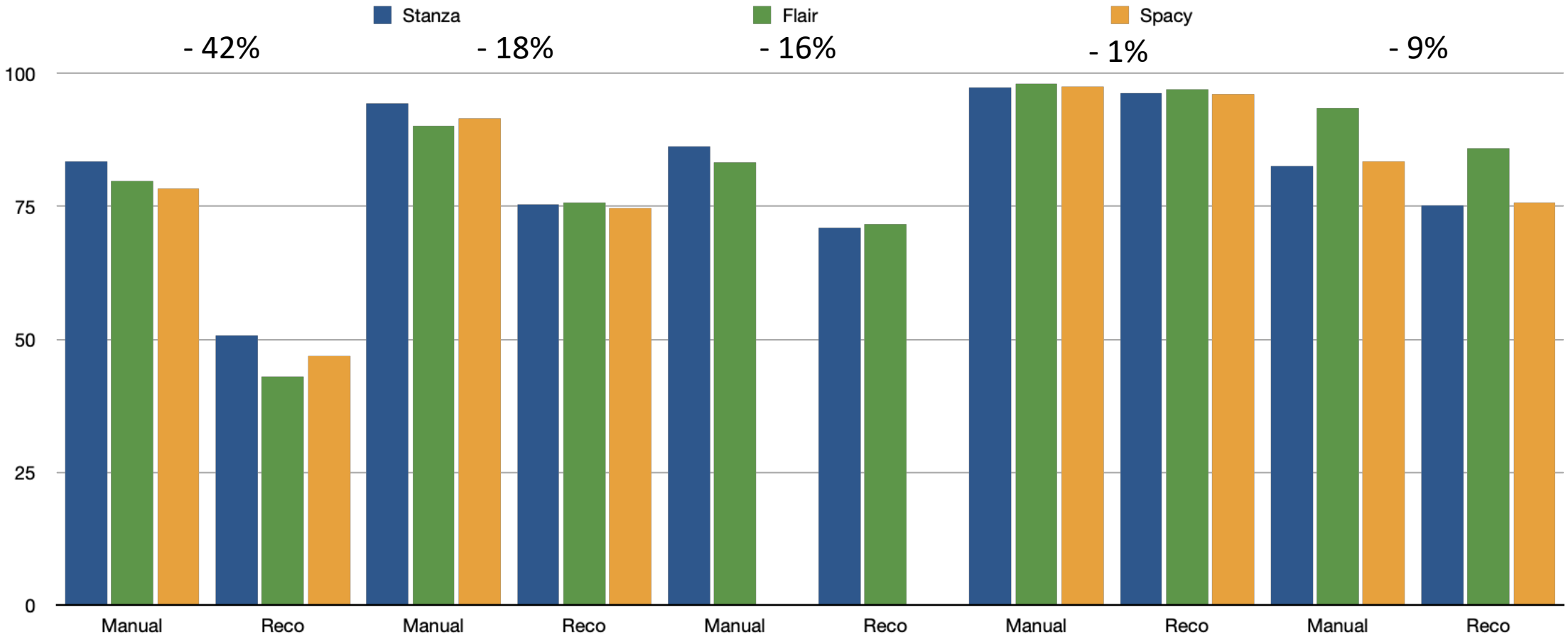


Language model Ngrams of BPE

	HOME				Balsac	Esposalles
AP@0.75	48.57				91.13	–
	Kaldi HOME (multilingual)				**Kaldi Balsac**	**Kaldi Esposalles**
	Czech	German	Latin	all	French	Catalan
CER	8.70	7.48	10.37	8.93	6.41	1.32
WER	29.71	26.40	35.59	29.26	17.41	3.51

TEKLIA

Results on automatic transcriptions



HOME-CZECH
CER: 8.7%

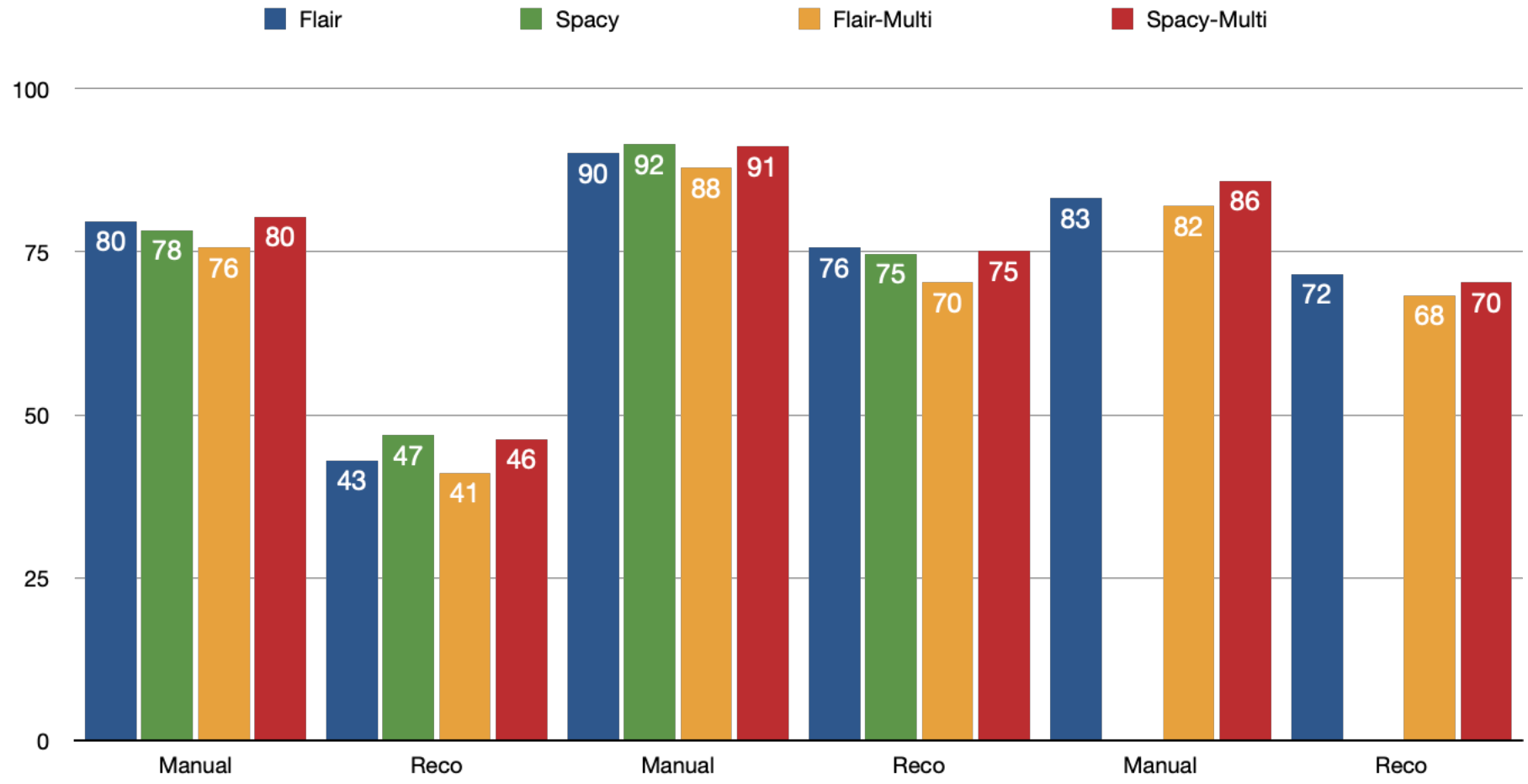
HOME-GERMAN
CER: 7.5%

HOME-LATIN
CER: 10.4%

ESPOSALLES
CER: 1.3%

BALSAC
CER: 6.4%

Results on autom. transc. with Multi-lingual NER



HOME-CZECH
CER: 8.7%

HOME-GERMAN
CER: 7.5%

HOME-LATIN
CER: 10.4%

Comparison to custom NER systems

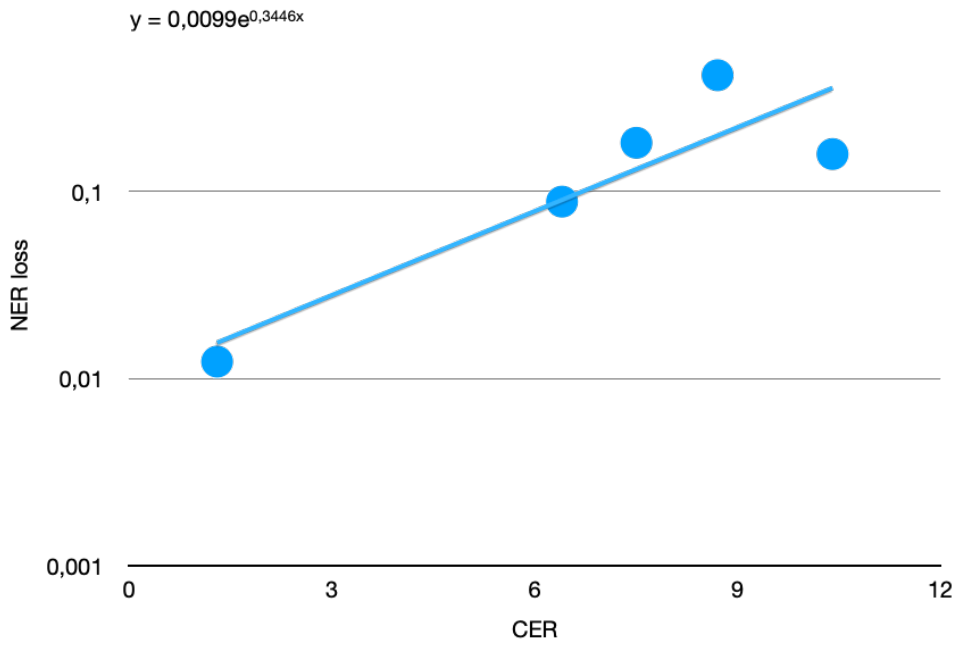
Esposalles

System	CER	NER F-score
Naver Labs	5%	0.95
CITlab-ARGUS-2	unk	0.92
Spacy	0%	0.98
	1.32%	0.96
Flair	0%	0.98
	1.32%	**0.97**
Stanza	0%	0.97
	1.32%	0.96

Conclusions

- Standard libraries are competitive
- No clear winner : choose one or test them all
- Multi-lingual models are a good approach when data is sparse for each language, both for HTR and NER
- Exponential relation CER-NER loss

- Invest your time/budget on line detection or HTR



Thank you - Merci - Gratias
tibi valde - Děkuji - Danke

TEKLIA | www.teklia.com

Christopher Kermorvant
kermorvant@teklia.com