



Corpora with Part-of-Speech Annotations for Three Regional Languages of France: Alsatian, Occitan and Picard

Delphine Bernhard, Anne-Laure Ligozat, Fanny Martin, Myriam Bras,
Pierre Magistry, Marianne Vergez-Couret, Lucie Steiblé, Pascale Erhart,
Nabil Hathout, Dominique Huck, Christophe Rey, Philippe Reynés,
Sophie Rosset, Jean Sibille, Thomas Lavergne
LiLPa, LIMSI, HM, CLLE, Queen's University, LT2D

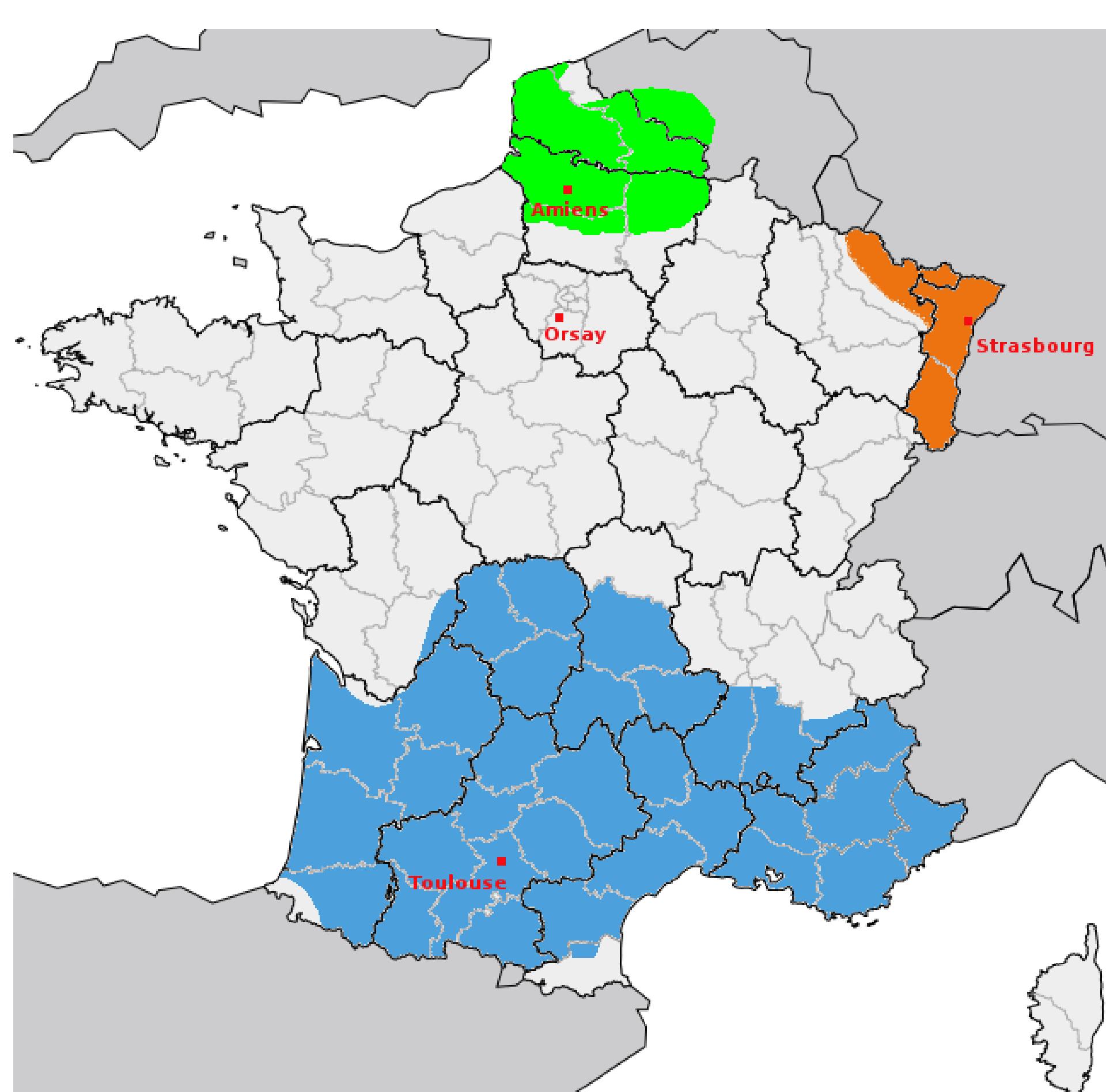
Context: the RESTAURE project

- Goal = provide computational resources and processing tools for three regional languages of France: Alsatian, Occitan and Picard
- Languages with no official status in France

Introduction

- Challenges:
 - lack of comprehensive grammatical descriptions, encompassing all the dialectal variants
 - tokenization issues
 - use of POS tagsets and taggers developed for closely related languages?
- Close cooperation between linguists and NLP specialists
- Parallel and collaborative work on three different languages facing similar challenges.

Languages



Additional Information & Download

- <http://restaure.unistra.fr/>
- <https://zenodo.org/communities/restaure/>

Common tagset

Tag	Full name
ADJ	adjective
ADP	adposition
ADP+DET	preposition-determiner contraction
ADV	adverb
AUX	auxiliary
CCONJ	coordinating conj.
DET	determiner
EPE	epenthesis
INTJ	interjection
MOD	modal verb
NOUN	common noun
NUM	numeral
PART	particle
PRON	pronoun
PROPN	proper noun
PUNCT	punctuation
SCONJ	subordinating conj.
SYM	symbol
VERB	verb
X	other

- Based on the POS tags defined in the Universal Dependencies project [2]
- Customised tagsets for each language, with mappings to the common tagset.

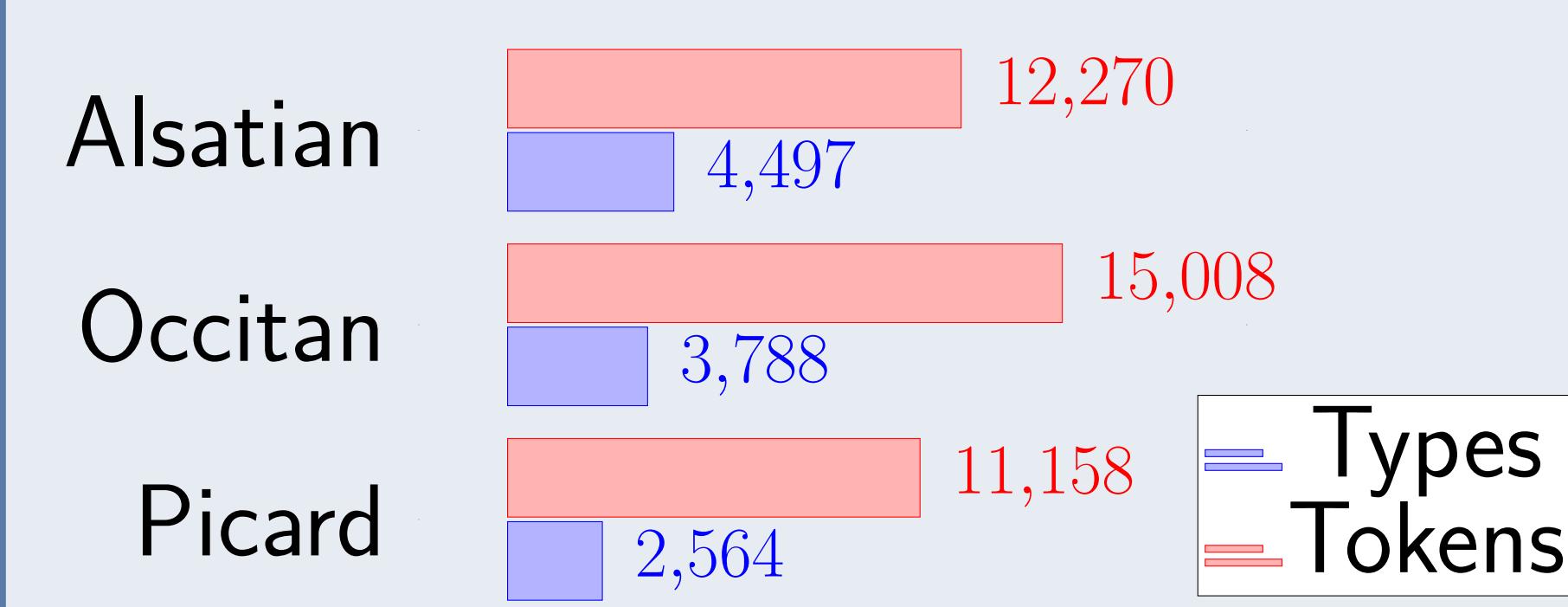
Methodology

- Design of the annotation guidelines
 - Alsatian: Universal POS tags + STTS tags for German + Alsatian grammars
 - Occitan: lexicon of inflected forms LOFLOC + GRACE tagset
 - Picard: Universal POS tags + French Treebank annotation guidelines
- Corpus selection:
 - texts with rights to distribute
- Corpus pre-processing:
 - tokenization with custom tokenizers for Alsatian and Occitan, manual for Picard
 - pre-annotation for Alsatian and Occitan
- Manual annotation:
 - POS tags, lemmas and potential additional information
 - several annotators
 - use of the AnaLog tool [1] for Alsatian and Occitan
 - IAA > .9 except for early annotation
- Manual and automatic final verifications

Acknowledgements

This work was supported by the French "Agence Nationale de la Recherche" (ANR) through the RESTAURE project (no.: ANR-14-CE24-0003). We also thank all the annotators.

Results: Reference corpora



Variation in: genre, dialect, publication dates.

Example annotations

Alsatian:

Token	Lemma	Tag	English
Spàrichle	Spàrichel	NOUN	asparagus
ìn	ìn	ADP	in
e	e	DET	a
Sibb	Sibb	NOUN	sieve
üss	üss	ADP	of
Metàll	Metàll	NOUN	metal
màche	màche	VERB	put

Occitan:

Token	Lemma	T1	T2	English
Los	lo	D	Da	the
cavals	caval	N	Nc	horses
èran	èsser	V	Vm	were
luènh	luènh	R	Rg	far away
.	.	F	.	.

Picard:

Token	Tag	Lemma	English
I	PRONPERS	i	he
avot	VERBCONJ	avoir	had
fauqu'	VERBPP	fauquer	cut
chés	DET	euch	the
projecteurs	NOUN	projecteur	spotlights
qu'	PRONREL	qui	that
is	PRONPERS	i	they
illuminotent	VERBCONJ	illuminoter	lit
eul	DET	euch	the
fosse	NOUN	fosse	pit
.	PUNCT	.	.

References

- [1] M.-H. Lay and B. Pincemin.
Pour une exploration humaniste des textes: AnaLog.
In Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles, pages 1045–1056, Sapienza University of Rome, 2010.
- [2] J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajic, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman.
Universal dependencies v1: A multilingual treebank collection.
In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), may 2016.