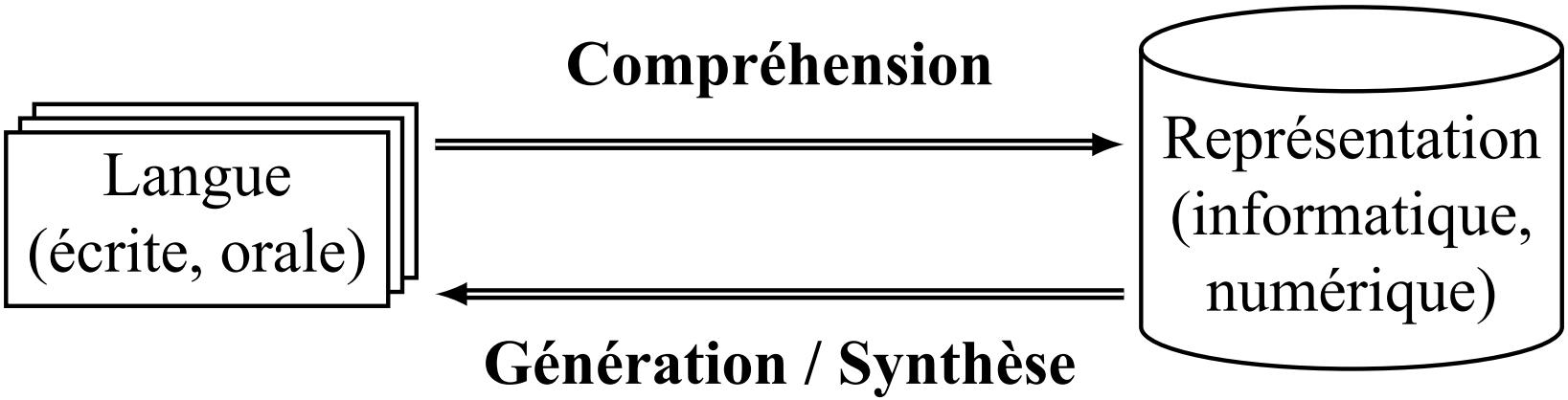
Traiter et désambiguiser les langues
Reconnaissances & Résolutions

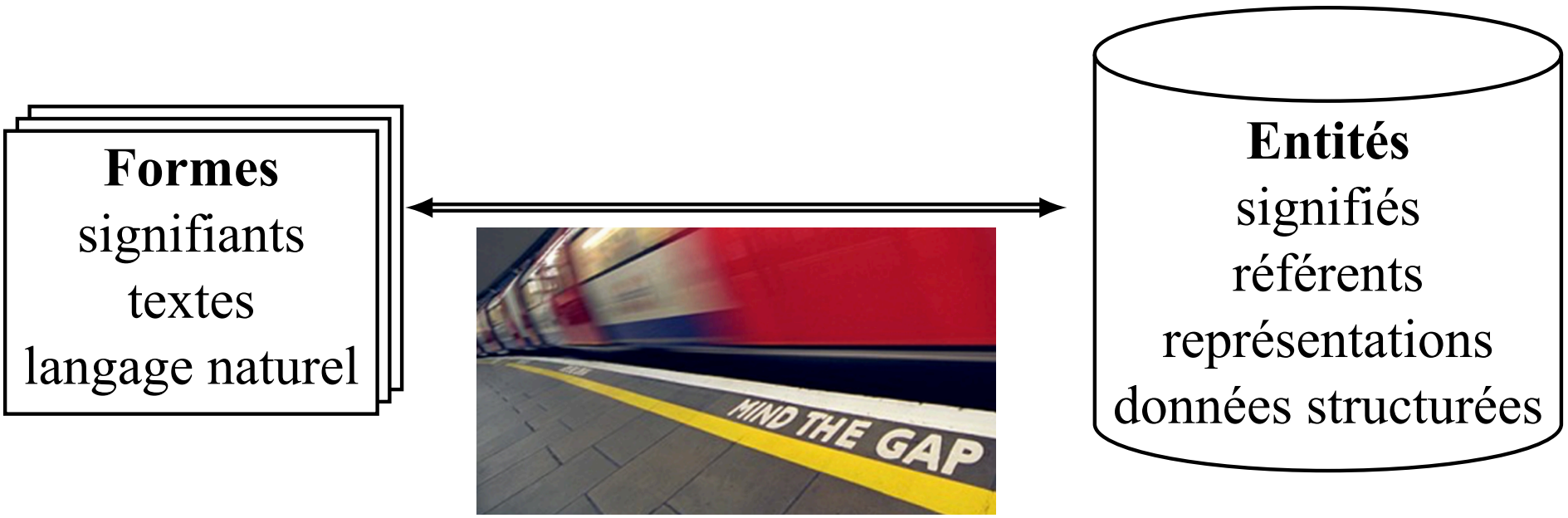
Damien Nouvel

Processus : génération vs compréhension



Plan

Du signal aux sens



- Faire le **lien** entre **signal** et **représentations** ?
- **Niveaux** d'analyse : caractères, mots, énoncés, documents ?
- Exploiter et tenir compte du **contexte** ?

Reconnaissons que ça n'est pas évident

« sss, l rgrdt l **pch** ps jst dvnt l …

Reconnaissons que ça n'est pas évident

« sss, l rgrdt l **pch** ps jst dvnt l …
l'nj tt mprtnt, t l sst d' vlr l rsq …

Reconnaissons que ça n'est pas évident

« sss, l rgrdt l **pch** ps jst dvnt l …
l'nj tt mprtnt, t l sst d' vlr l rsq …
l s'gsst d n ps s trmpr, ctt fs-c …

Reconnaissons que ça n'est pas évident

« sss, l rgrdt l **pch** ps jst dvnt l …
l'nj tt mprtnt, t l sst d' vlr l rsq …
l s'gsst d n ps s trmpr, ctt fs-c …
ls trs tnt ssmbls tr d l, n slnc …

Reconnaissons que ça n'est pas évident

« sss, l rgrdt l **pch** ps jst dvnt l …
l'nj tt mprtnt, t l sst d' vlr l rsq …
l s'gsst d n ps s trmpr, ctt fs-c …
ls trs tnt ssmbls tr d l, n slnc …
ps, fnlmnt, l s dcd gr ! »

Reconnaissons que ça n'est pas évident

« sss, l rgrdt l **pch** ps jst dvnt l …
l'nj tt mprtnt, t l sst d' vlr l rsq …
l s'gsst d n ps s trmpr, ctt fs-c …
ls trs tnt ssmbls tr d l, n slnc …
ps, fnlmnt, l s dcd gr ! »

⇒ Reconstruire un texte est difficile …

Désambiguiser or not désambiguiser

« Assis, il regardait la **pioche** posée juste devant lui

Désambiguiser or not désambiguiser

« Assis, il regardait la **pioche** posée juste devant lui
l'enjeu était important, et il essayait d'évaluer le risque

Désambiguiser or not désambiguiser

« Assis, il regardait la **pioche** posée juste devant lui
l'enjeu était important, et il essayait d'évaluer le risque
il s'agissait de ne pas se tromper, cette fois-ci

Désambiguiser or not désambiguiser

« Assis, il regardait la **pioche** posée juste devant lui
l'enjeu était important, et il essayait d'évaluer le risque
il s'agissait de ne pas se tromper, cette fois-ci
les autres étaient assemblés autour de lui, en silence

Désambiguiser or not désambiguiser

« Assis, il regardait la **pioche** posée juste devant lui
l'enjeu était important, et il essayait d'évaluer le risque
il s'agissait de ne pas se tromper, cette fois-ci
les autres étaient assemblés autour de lui, en silence
puis, finalement, il se décida ! »

Désambiguiser or not désambiguiser

« Assis, il regardait la **pioche** posée juste devant lui
l'enjeu était important, et il essayait d'évaluer le risque
il s'agissait de ne pas se tromper, cette fois-ci
les autres étaient assemblés autour de lui, en silence
puis, finalement, il se décida ! »

$\Rightarrow$ On ne peut toujours comprendre avec 100% de certitude …

De l'ambiguïté

« MAGNIFIQUE PORTE ! »

⇒ Ce texte ne semble guère ambigu …

De l'ambiguïté

« MAGNIFIQUE PORTE ! »

⇒ Ce texte ne semble guère ambigu …

« Magnifique porté ! »

⇒ Désambiguisation par le lexique (système d'écriture)

De l'ambiguïté

« MAGNIFIQUE PORTE ! »

⇒ Ce texte ne semble guère ambigu …

« Magnifique porté ! »

⇒ Désambiguisation par le lexique (système d'écriture)

« CE PATINEUR A FAIT UN MAGNIFIQUE PORTE ! »

⇒ Désambiguisation contextuelle

Cadre général

- La désambiguisation pour la
  - **Compréhension** (communication)
  - **Représentation** (mentale, encodée, partielle)

$\Rightarrow$ Importance des informations **contextuelles** (auteur, date, etc.)

Cadre général

- La désambiguisation pour la
  - **Compréhension** (communication)
  - **Représentation** (mentale, encodée, partielle)

⇒ Importance des informations **contextuelles** (auteur, date, etc.)

- Traitement automatique des Langues …
  - Quelle **langue** (naturelle)
  - Quelle **modalité** (oral / écrit / signé / etc.)
  - Des **variations** : diachronie, domaines de spécalités, écritures

Cadre général

- La désambiguisation pour la
  - **Compréhension** (communication)
  - **Représentation** (mentale, encodée, partielle)

⇒ Importance des informations **contextuelles** (auteur, date, etc.)

- Traitement automatique des Langues …
  - Quelle **langue** (naturelle)
  - Quelle **modalité** (oral / écrit / signé / etc.)
  - Des **variations** : diachronie, domaines de spécalités, écritures
- Exploitation de **ressources**

⇒ Corpus, lexiques (morphologie), grammaires (syntaxe)

⇒ Disparité selon les langues (multilinguisme / plurilinguisme)

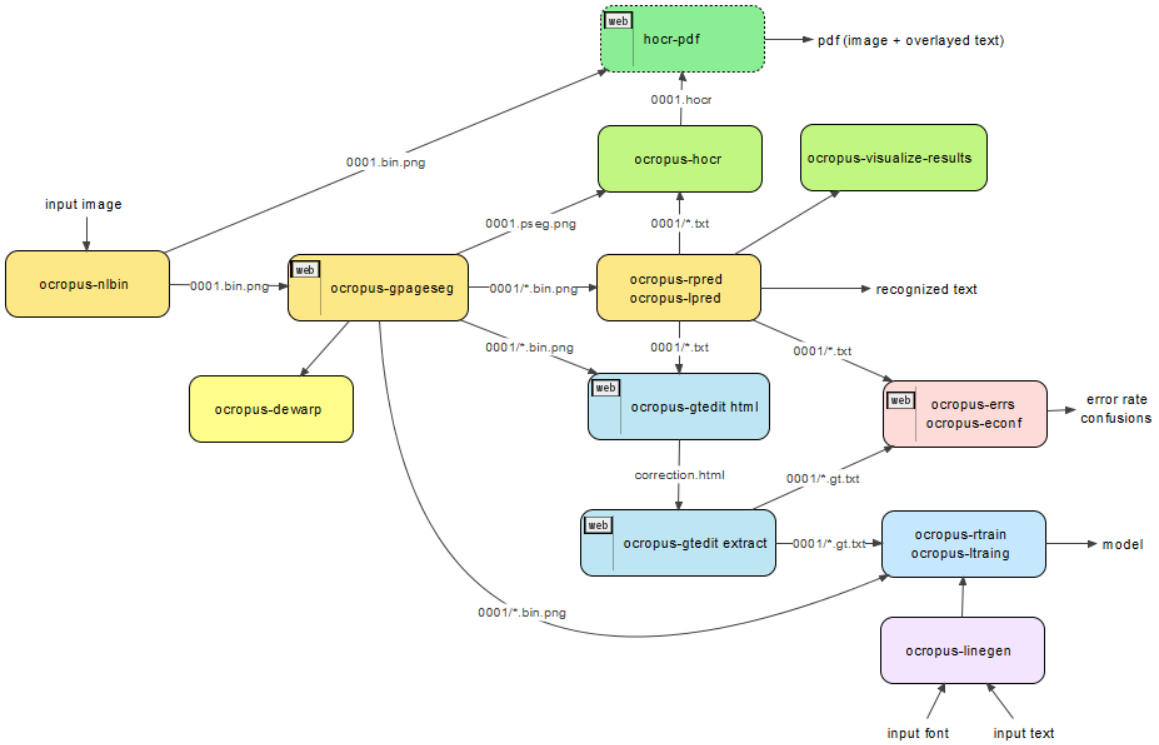
Plan

Objectifs généraux

- Projet CNES / ERTIM (deux ingénieurs)
  - Acquisition, extraction, prétraitement des données
  - Reconnaissance à partir d'images

  $\Rightarrow$ Adaptation d'un OCR (OCRopy) libre et gratuit
  - Détection de « signaux faibles »

Architecture d'Ocropy

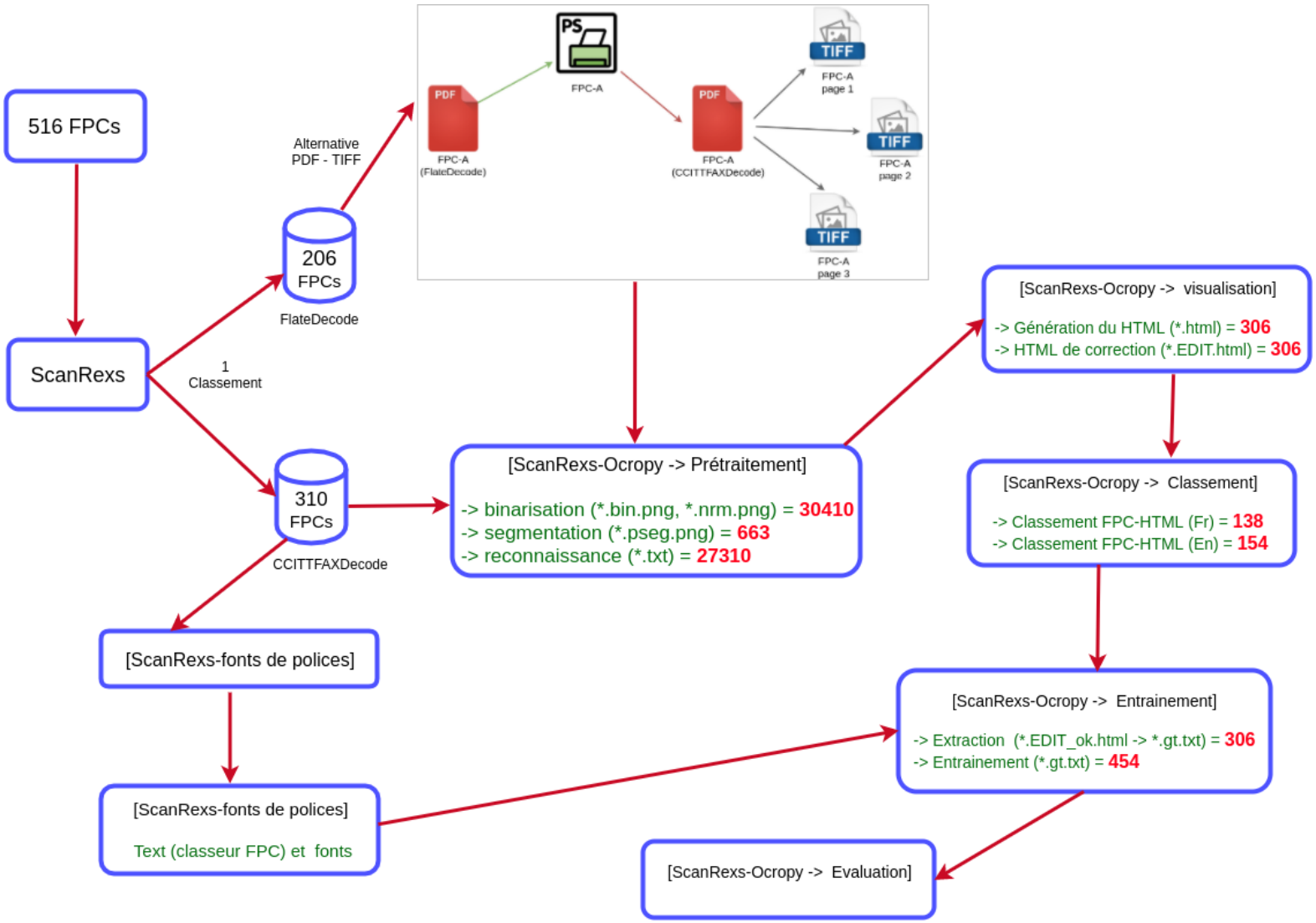


- Comparaison de 3 OCRs (Springmann *et. al.*, 2014) :

$\Rightarrow$ Ocropy 81.66% / ABBYY 80.57% / Tesseract 78.77%)

- Textes historiques (finnois) (Drobac *et. al.* 2017) : 93% – 95.21%

Chaîne de traitement adoptée : quelques chiffres



Méthodologie

- Reconnaissance des caractères sur les images
- Génération des fichiers d'édition (HTML)
- Correction des fichiers d'édition (HTML)
- Génération de *Ground Truth*
- Apprentissage des modèles (IA, données *train*)
- Evaluation des modèles (données test)
- Choix du meilleur modèle

Constitution des corpus

- Corpus d'apprentissage : 44 FPCs corrigées + Ocropy default

Classe	Taux OOV	initial	default
A	0	0	0
B	10	0	0
C	20	16	25
D	30	60	88
E	40	99	74
F	50	48	48
G	60	30	46
H	70	14	4
I	80	5	0
J	90	23	13

Correction des documents numérisés

- Correction **manuelle** : indispensable, coûteux en temps
- Estimation des temps de correction

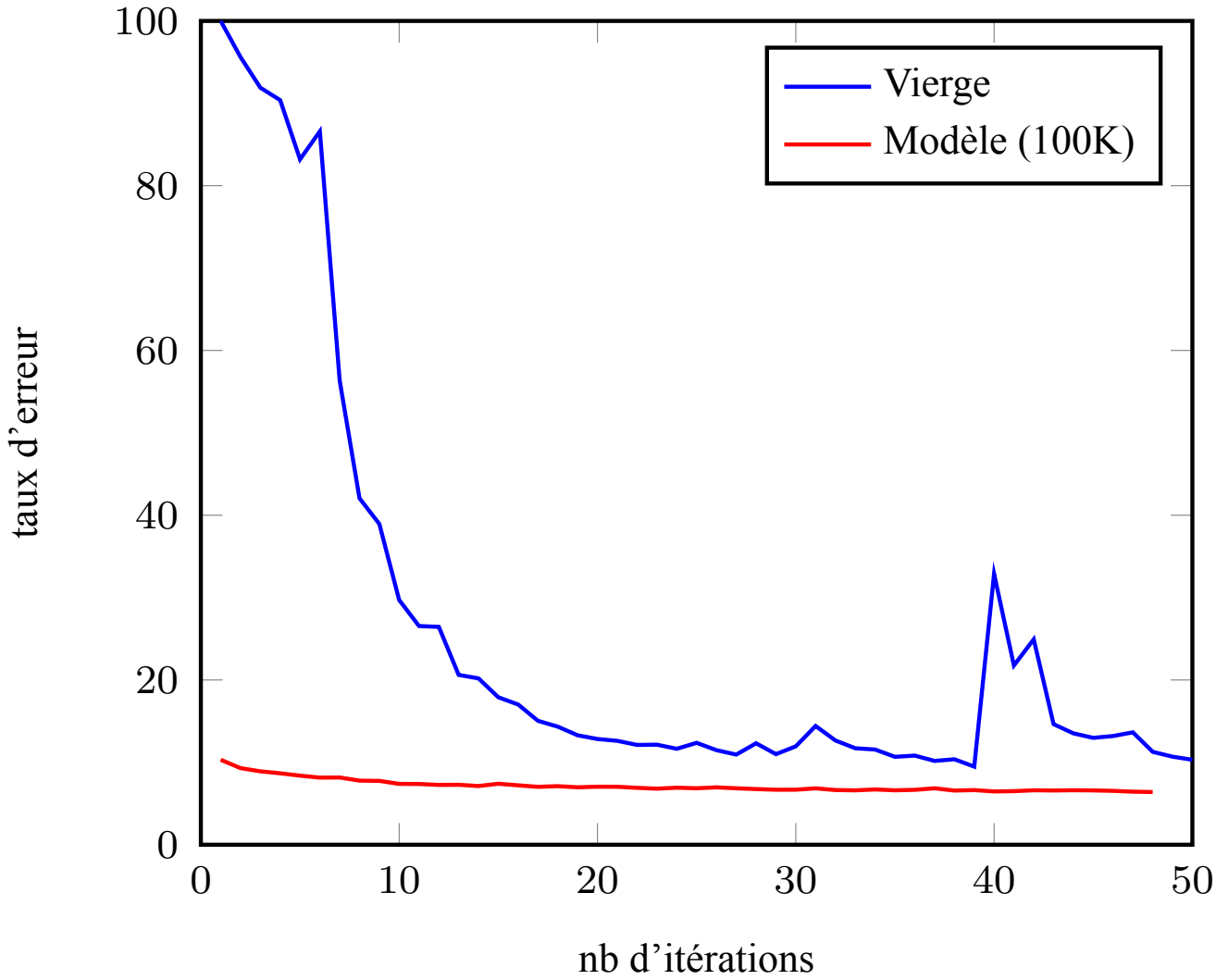
html	1-2 p.	1-5 p.	5 p.+
A	5mn	10mn	15mn+
B	10mn	20mn	30 mn+
C	15mn	25mn	35 mn+
D	15-20mn	25-40mn	60mn+

Données et paramètres d'apprentissage

- Apprentissage automatique : adaptation ocropy aux données
- Itération des modèles et calcul de performance (taux d'erreurs)
- Utilisation de fiches corrigées (*ground truth*)

Récapitulatif des entraînements réalisés	
nbre de passes	18
données d'entraînement	44 à 109 FPC
données de test	10 à 15 FPC
nbre de modèles générés	1334 (980+305+49) modèles
nbre d'itérations atteint	plus de 340000
meilleur modèle obtenu	CER 4, 271%
nbre d'itérations meilleur modèle	290000

Évaluations selon les itérations



Interprétation des résultats

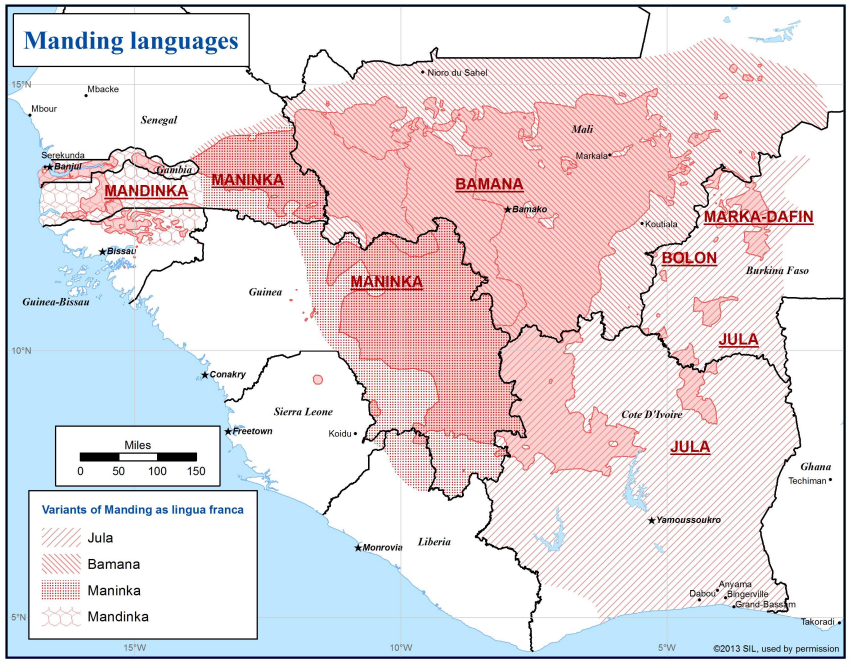
- Quelques éléments d'interprétation
  - Le nombre d'itération améliore les performances
  - Moins visible avec le modèle par défaut mais $10.315 \rightarrow 6.398$
  - Impact
    - Du travail de correction
    - De la quantité de données d'entraînement
    - De l'ajout des codecs de FPC
    - Impact des modèles pré-entraînés

Quelques conclusions

- Numérisation des fiches FPC
  - Développement et adaptation dun OCR
  - Annotation de fiches pour son entraînement
  - Évaluations pour valider sa fiabilité …mais à confirmer !
  - Livraison du logiciel et des données en cours
- Travaux en détection de signaux faibles
  - Difficulté de détecter automatiquement la gravité
  - Extraction de vocabulaire selon
    - La gravité de la fiche (risque)
    - La date de la fiche (campagnes)

Plan

Contexte



- Parlée principalement au Mali (diglossie : français, 20%)
- Ou « bamanaka »
- 4M de locuteurs (10M en 2 ème langue)
- Véhiculaire, tradition orale
- Macro-langue mandingue (avec dioula, malinké, etc.)

Quelques considérations linguistiques

- Prononciation : 7 voyelles, 20 consonnes, 3 tons

Quelques considérations linguistiques

- Prononciation : 7 voyelles, 20 consonnes, 3 tons
- Alphabet
  - Langues mandingues : alphabet n'ko (1950, Unicode 5.0, rtl)
  - Bambara : alphabet latin
  - Depuis 1982 : ɛ (U+025B/U+03B5), ɔ, ŋ, ɲ

$\Rightarrow$ Langues peu standardisées

Quelques considérations linguistiques

- Prononciation : 7 voyelles, 20 consonnes, 3 tons
- Alphabet
  - Langues mandingues : alphabet n'ko (1950, Unicode 5.0, rtl)
  - Bambara : alphabet latin
  - Depuis 1982 : ɛ (U+025B/U+03B5), ɔ, ŋ, ɲ

$\Rightarrow$ Langues peu standardisées

- Grammaire
  - Type : S AUX O V X, tonale
  - Pas de genre grammatical
  - Pas de conjugaison (marques prédicatives AUX)
  - Peu de flexion (-w : pluriel)

Le corpus bambara de référence

- Collecte de textes en bambara
  - Publiés (périodiques, littérature) ou non (lettres, trans.)
  - Normalisation des textes (orthographe, tons, etc.) ?
  - Textes en ligne `http://cormand.huma-num.fr/biblio/`

⇒ Volume : 2,3M mots

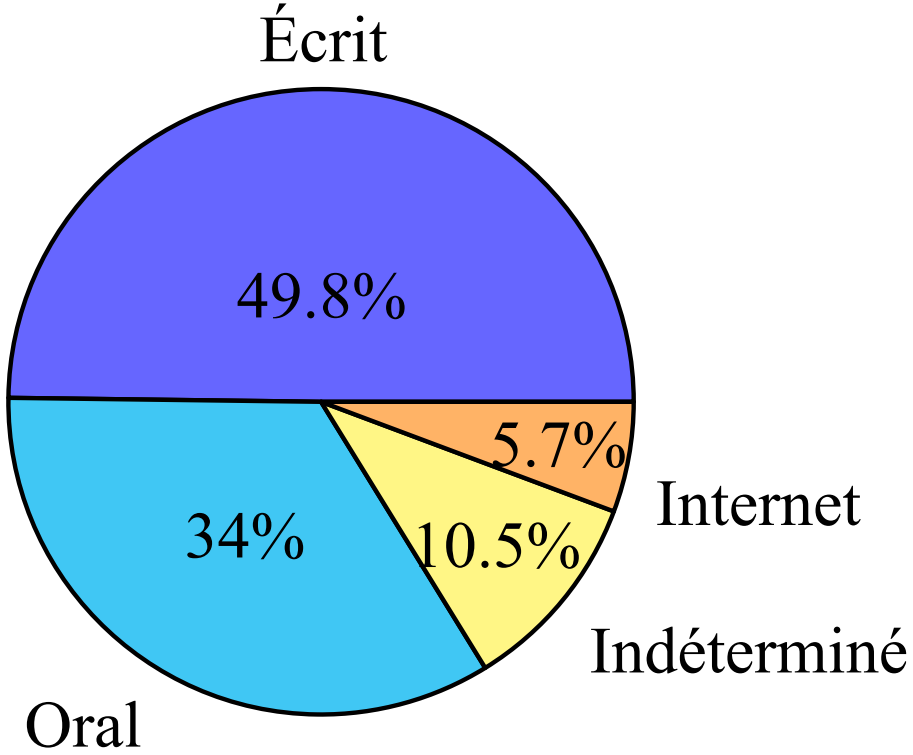
- Utilisation essentiellement linguistique
  - Apprentissage de la langue
  - Études linguistiques sur corpus
  - Annotation (POS, lemmes, morphologie, gloses)

⇒ Labex EFL, axe 6 : ressources linguistiques

⇒ Site internet : `http://cormand.huma-num.fr/`

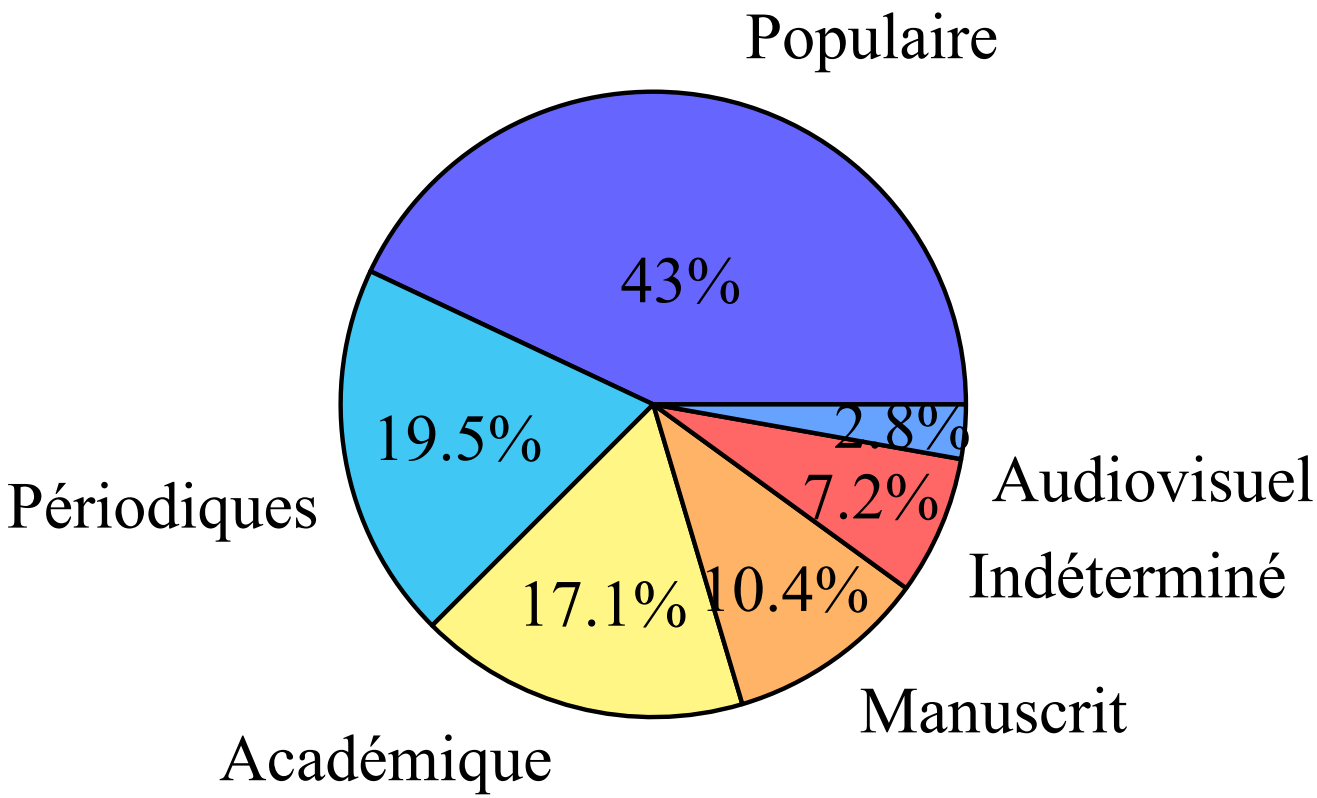
⇒ Modélisation linguistique informatisée (SketchEngine)

Médiums du corpus



⇒ Prédominance de l'écrit, mais parfois issu de l'oral (contes)

Sources du corpus



Annotation du corpus

- Utilisation de **Daba** (Maslinsky, 2014)
  - Tokenisation
  - Recherche dans les dictionnaires
  - Analyses morphologiques

Annotation du corpus

- Utilisation de **Daba** (Maslinsky, 2014)
  - Tokenisation
  - Recherche dans les dictionnaires
  - Analyses morphologiques

$\Rightarrow$ Pré-annotation automatique et ambiguë

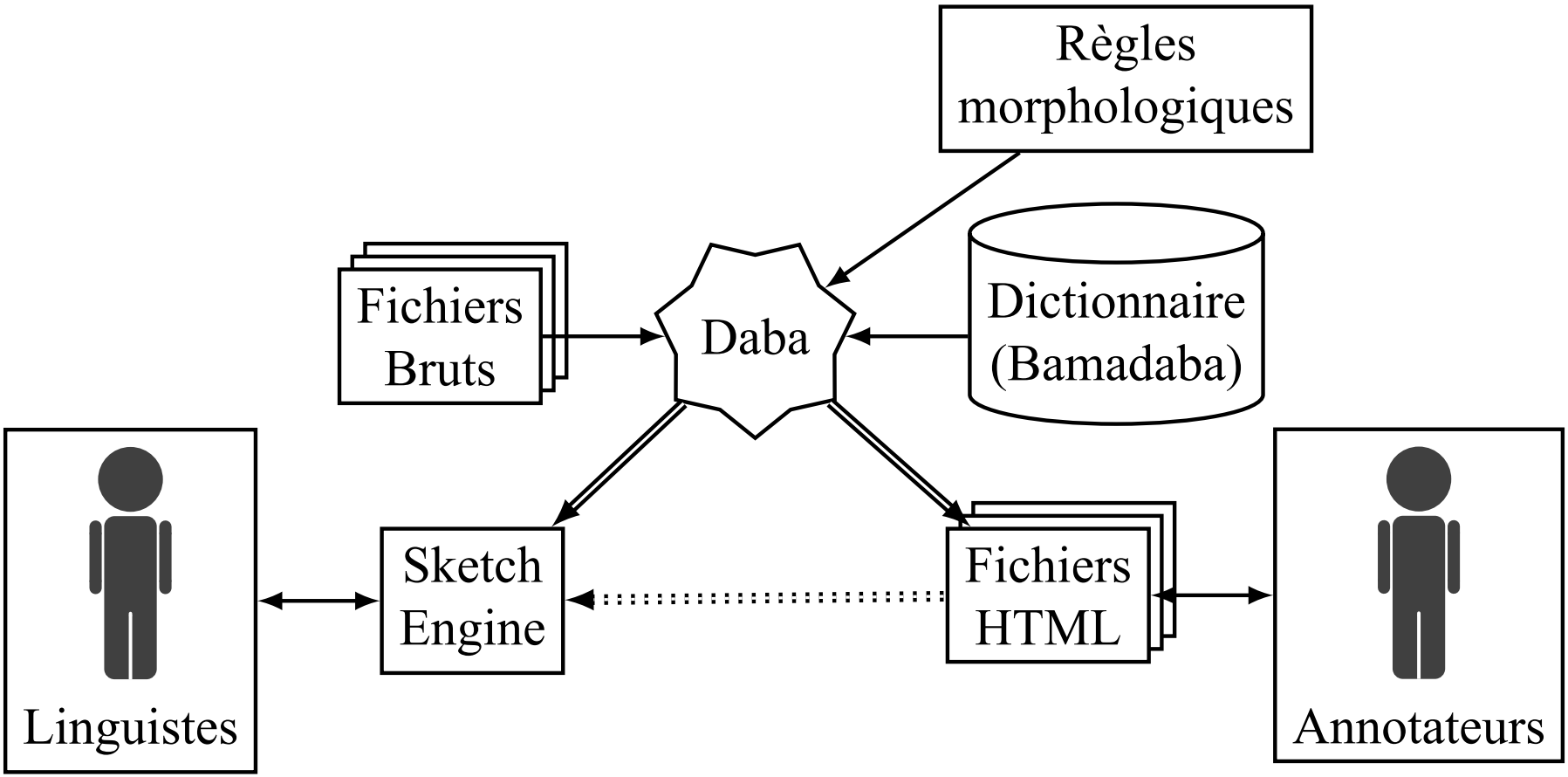
Annotation du corpus

- Utilisation de **Daba** (Maslinsky, 2014)
  - Tokenisation
  - Recherche dans les dictionnaires
  - Analyses morphologiques

$\Rightarrow$ Pré-annotation automatique et ambiguë

- Annotation manuelle
  - Peu de moyens (profs, étudiants, bénévoles)
  - Validation humaine
  - Correction / normalisation

Schéma de fonctionnement



Objectifs de MANTAL

- Traitement des langues MANdingues avec des outils TAL
  - Normalisation / uniformisation
  - Parties du discours
  - Tonalisation
  - (Entités nommées)
  - (Malinké)

$\Rightarrow$ Exploitation des données annotées (apprentissage)

Objectifs de MANTAL

- Traitement des langues MANdingues avec des outils TAL
  - Normalisation / uniformisation
  - Parties du discours
  - Tonalisation
  - (Entités nommées)
  - (Malinké)

⇒ Exploitation des données annotées (apprentissage)

- Cadre général
  - Collaboration LLACAN / ERTIM
  - Durée : 3 ans
  - Budget interne Inalco (stages, prestations, missions)

⇒ Interactions avec les linguistes

Procédures d'annotation

- Stage M2 « Annotation automatique en POS pour le Bambara »

$\Rightarrow$ D. Auffret

- Plusieurs niveaux de traitement

Procédures d'annotation

- Stage M2 « Annotation automatique en POS pour le Bambara »

⇒ D. Auffret

- Plusieurs niveaux de traitement
  - Fichier source
  - Recherche des ponctuations, nombres et noms propres
  - Utilisation de Daba
    - Recherche dans les dictionnaires
    - Analyse morphologique

    ⇒ Automatique, sortie ambiguë
  - Annotation par les linguistes

  ⇒ Ambiguïtés résiduelles

⇒ Plusieurs versions de fichiers à synchroniser

Fichier source

<h>Diɲɛ Yaalala</h>
Nsiirin, nsiirin. N y'a bila den dɔ le kan.
Den nin ye sira dali a facɛ fɛ, k'a b'a fɛ ka taga diɲɛ yaala ka dɔ fara a hakili kan. A facɛ ye sira d'a ma a ka taga yaala. A tagara yaala kosɛbɛ.
A fɔlɔla ka taga ben sogosu dɔ ma. A yɔrɔ bɛɛ tolira, a ko : "E ! Ala bɛ se." Sogo nin wulila ka kum'a fɛ k'a kan'a fɔ ko Ala bɛ se, k'Ala ka se b'a ɲɛfɛ. "A tɛmɛna sogo nin na ka taga ɲɛfɛ, ka kɔlɔn saba ye. Fɔlɔ jalen bɛ, ji foyi t'a la. A filanan ji to kɔnɔ. A sabanan ji b'o kɔnɔ. A tɛmɛn'o la ka taga se cɛkɔrɔnin dɔ ma. Cɛ nin kɔrɔla kosɛbɛ. [...]

Fichier Daba

Nsiirin, nsiirin.

Nsiirin	,	nsiirin	.
nsíirin		nsíirin	
n		n	
conte		conte	

N y'a bila den dɔ le kan.

N	y'	a	bila	den	dɔ	le	kan	.
ń	y'	à	bìla	dén	dɔ́	le	kàn	
pers	pm	pers	v	n	dtm	prt	pp	
1SG	PFV.TR	3SG	mettre	enfant	certain	FOC	sur	

Fichier vertical

Token	Lemma	PdD	Glose	Compos.	Original	Tonal
nsiiri	nsiiri	n	conte		NSIIRI	nsiiri
naaninan	naaninan	ORD|adj	quatrième	naani	NAANINAN	náaninan
dinye	dunya|jinye |dunu-nye|dinye |jyen|diyen	n	monde		Diɲɛ	díɲɛ
yaalala	yaalala	AG.PRM|n		yaala	Yaalala	yáalala
nsiirin	nsiiri|nsiirin	n	conte		Nsiirin	nsíirin
,	,	c	,	,	,	,
nsiirin	nsiiri|nsiirin	n	conte		nsiirin	nsíirin
.	.	c	.	.	.	.
n	n	pers	1SG		N	ń
y'	ye|y'	pm	PFV.TR		y'	y'
a	a	pers	3SG		a	à
bila	bil'|bla|bila	v	mettre		bila	bìla
den	den	n	enfant		den	dén
do	do	dtm	certain		dɔ	dɔ́
le	le	prt	FOC		le	le
kan	kan	pp	sur		kan	kàn
.	.	c	.	.	.	.

Volumétrie des ressources

- Corpus

Corpus	Balises	Ponctuations	Formes (distinctes)
Brut	412K	383K	2 321K (68K)
Désambiguïsé	104K	71K	426K (19K)

- Dictionnaires (disponibles en ligne)

Dictionnaire	Description	Entrées	Ambiguïté
bamadaba	Dictionnaire principal	11K	1,137
enciclop	Notions encyclopédiques	29	1
jamuw	Noms claniques	375	1,001
togow	Prénoms	496	1
yorow	Toponymes	299	1

Jeu d'étiquettes sur le corpus désambiguisé

Code	Partie du discours	Quantité
n	nom	82K
c	ponctuation	66K
pers	pronom personnel	54K
v	verbe	51K
pm	marque prédicative	41K
pp	postposition	34K
conj	conjonction	21K
cop	copule	18K
n.prop	nom propre	12K
dtm	déterminatif	12K
prn	pronom (non-personnel)	10K
prt	particule	10K

num	numératif	6K
adj	adjectif	4K
ptcp	participe	4K
intj	interjection	2K
adv	adverbe	2K
vq	verbe qualitatif	1K
onomat	onomatopée	102
adv.p	adverbe préverbal	26
conv.n	converbe nu	24
mrph	morphème	13

$\Rightarrow$ Relativement standard

Apprentissage : algorithme et données

- Méthodologie : apprentissage (Wapiti, Lavergne, 2010)
- Ingénierie de traits (features)
  - Préfixe de 2 et 3 caractères
  - Suffixe de 2 et 3 caractères
  - Version transformé du mot par expression régulière (ANP)
  - Version non-tonalisée du mot
  - Taille du mot
  - Étiquettes possibles dans les dictionnaires

Apprentissage : algorithme et données

- Méthodologie : apprentissage (Wapiti, Lavergne, 2010)
- Ingénierie de traits (features)
  - Préfixe de 2 et 3 caractères
  - Suffixe de 2 et 3 caractères
  - Version transformé du mot par expression régulière (ANP)
  - Version non-tonalisée du mot
  - Taille du mot
  - Étiquettes possibles dans les dictionnaires

Score	Base	Non-ton.	Préf.	Suf.	Dico	ANP	Taille	Tous
Mot U	86.14	86.36	89.09	89.46	89.87	89.41	89.14	90.65
Phrase U	18.70	19.09	24.38	25.60	25.62	25.18	23.82	29.04
Mot B	85.94	91.27	91.69	91.91	91.50	86.84	88.37	**94.22**
Phrase B	13.14	35.35	36.39	37.51	33.94	15.20	21.57	**47.90**

Comparaison avec TreeTagger

- Tests TreeTagger
  - Entraînement : 90% du corpus, validation croisée
  - Pas de ressource additionnelle
  - Configuration par défaut

Comparaison avec TreeTagger

- Tests TreeTagger
  - Entraînement : 90% du corpus, validation croisée
  - Pas de ressource additionnelle
  - Configuration par défaut

Outil	**Score**
Baseline	22%
Majorité	82,06%
TreeTagger	93,50
Wapiti	94.22

Utilisation des tons

- Caractéristiques
  - Trois marques tonales : ̀, ́, ̌ (caron, hatchek)

  $\Rightarrow$ Change le sens du mot
  - Exemples :
    - bá = maman / bà = chèvre
    - tùgu = bras / túgu = fermer
    - tà = prendre, porter / tá = feu, propriété
- Les tons sont peu souvent marqués à l'écrit
- Essentiellement en 1 ère syllabe

Utilisation des tons

- Caractéristiques
  - Trois marques tonales : ̀, ́, ̌ (caron, hatchek)

  $\Rightarrow$ Change le sens du mot
  - Exemples :
    - bá = maman / bà = chèvre
    - tùgu = bras / túgu = fermer
    - tà = prendre, porter / tá = feu, propriété
- Les tons sont peu souvent marqués à l'écrit
- Essentiellement en 1 ère syllabe

$\Rightarrow$ La présence de tons aide pour la morpho-syntaxe

$\Rightarrow$ Détecter automatiquement les tons ?

Entropie des tons

- Méthodologie
  - Probabilités de tonalisations
  - Calcul d'entropie par mot

Entropie des tons

- Méthodologie
  - Probabilités de tonalisations
  - Calcul d'entropie par mot

Ent.	Tonalisations	Traduction
3.20	táatúmà : 1.0 , tàatúmá : 1.0 , táatúmá : 1.0 , táatùma : 9.0 , táatúma : 1.0 , tàatúmà : 1.0 , tàatúma : 1.0 , tàatùmà : 1.0 , tàatùmá : 1.0 , táatùmá : 1.0 , táatùmà : 2.0 , tàatuma : 1.0 , tàatùma : 1.0 , táatuma : 1.0	Départ (?)
2.50	bámànankɛ : 10.0 , bámànakɛ̀ : 16.0 , bámànankɛ̀ : 5.0 , bámànánkɛ : 6.0 , cɛ̀ : 16.0 , bámànanke : 1.0 , bámànan : 16.0 , bámanankɛ : 1.0	Bambara
2.25	súurun : 1.0 , súuru : 1.0 , sùruntu : 2.0 , súruntu : 1.0 , sùrundu : 1.0	Verser
2.02	cíyɛn : 9.0 , tìɲɛ : 11.0 , tíɲɛ : 2.0 , ciyén : 2.0 , cìyɛn : 11.0	Vérité
2.00	ɲɛ̀nen : 1.0 , ɲɛ́nen : 1.0 , ɲànen : 1.0 , ɲɛ̀ : 1.0	Regard
1.99	ɲɔ́gɔri : 7.0 , nwàri : 1.0 , ɲɔ̀gɔri : 1.0 , ɲɔ́ɔri : 5.0 , ɲúari : 5.0	Approcher
1.99	ɲɛ̀nen : 1.0 , ɲɛ́nen : 1.0 , ɲànen : 1.0 , ɲɛ̀ :? 1.0	?
1.99	ɲɔ́gɔri : 7.0 , nwàri : 1.0 , ɲɔ̀gɔri : 1.0 , ɲɔ́ɔri : 5.0 , ɲúari : 5.0	Salir
1.99	lé : 1.0 , lè : 4.0 , l̀è : 1.0 , le : 7.0 , dè : 4.0	(clan)
1.95	tàamasyɛnw : 2.0 , tàamashyɛn : 2.0 , táamashyɛnw : 1.0 , tàamashyɛnw : 2.0	Indiquer

Tonalisation automatique

- Stage M1 « Désambiguïsation lexicale des corpus bambara »

$\Rightarrow$ E. Mboning

- En moyenne deux tonalisations possible par mot non-tonalisé
- Apprentissage automatique (NLTK)
- Division train / test : 80% / 20%
- Essai de plusieurs algorithmes
  - CRF (peu de features) : 73%
  - Séquentiel (bigrammes, trigrammes, etc.) : 80%

Tonalisation automatique

- Stage M2 « Traitements TAL pour le bambara / maninka »

$\Rightarrow$ L. Liu, INALCO

$\Rightarrow$ Article IJNLP « A Bambara Tonalization System for WSD »

System Architecture

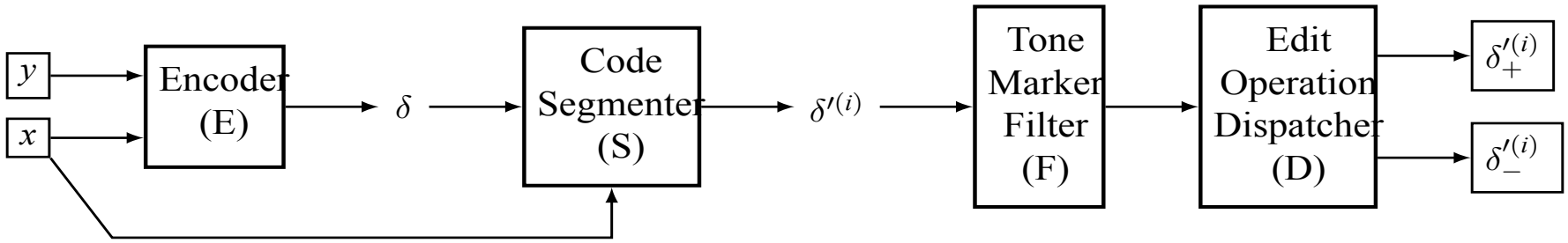


Figure – Block diagram for the proposed Bambara tonalization system at training stage

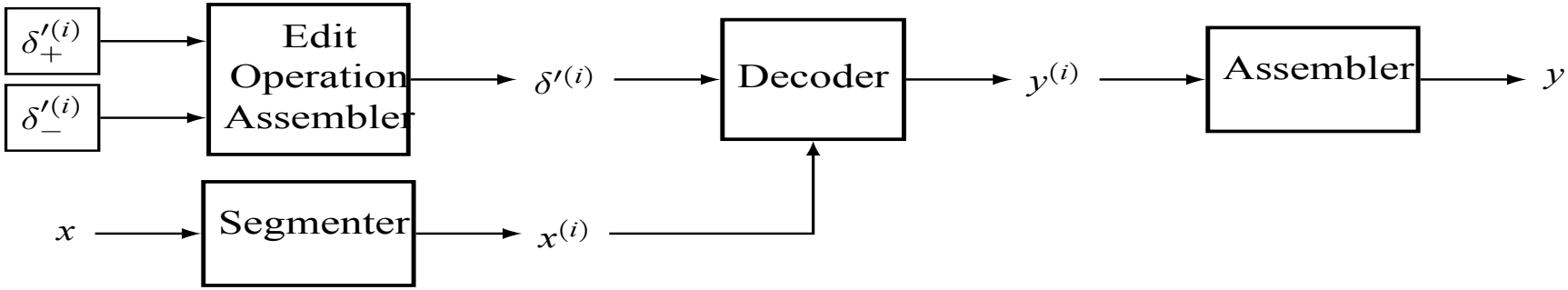


Figure – Block diagram for the proposed Bambara tonalization system at tonalization stage

Fundemental definitions

- Discrete random variables

$$X \longrightarrow \text{non-tonalized token} : \textit{kelen}$$
$$Y \longrightarrow \text{tonalized token} : \textit{kèlen}(\text{adj. same}), \textit{kèlén}(\text{intj. already})$$
$$\Delta \longrightarrow \text{differential code} : (+1, 2, \acute{}), (+1, 2, \acute{})(+1, 4, \grave{})$$

- Mappings

$$\Delta = E(Y; X) \longrightarrow \text{encoder function}$$
$$Y = D(\Delta; X) \longrightarrow \text{decoder function}$$
$$Y = D(E(Y; X); X)$$

- Predict differential code Δ, recovery Y from Δ by decoder D.

Experiment Result

- About half (52.35%) of tokens in BRC do not need tones

Sys. \ w	-1 (Syll.)	1	2	3	4	0
Majority vote	**0.843**					
S ∘ E	**0.923**	0.915	**0.922**	**0.922**	0.917	0.893
time	101.63	25.52	42.03	235.35	378.37	2683.72
D ∘ F ∘ S ∘ E	**0.923**	0.912	**0.923**	**0.923**	0.918	0.893
time	19.88	17.62	13.17	15.67	19.62	261.83

Table – Accuracy for the system trained

Experimental results

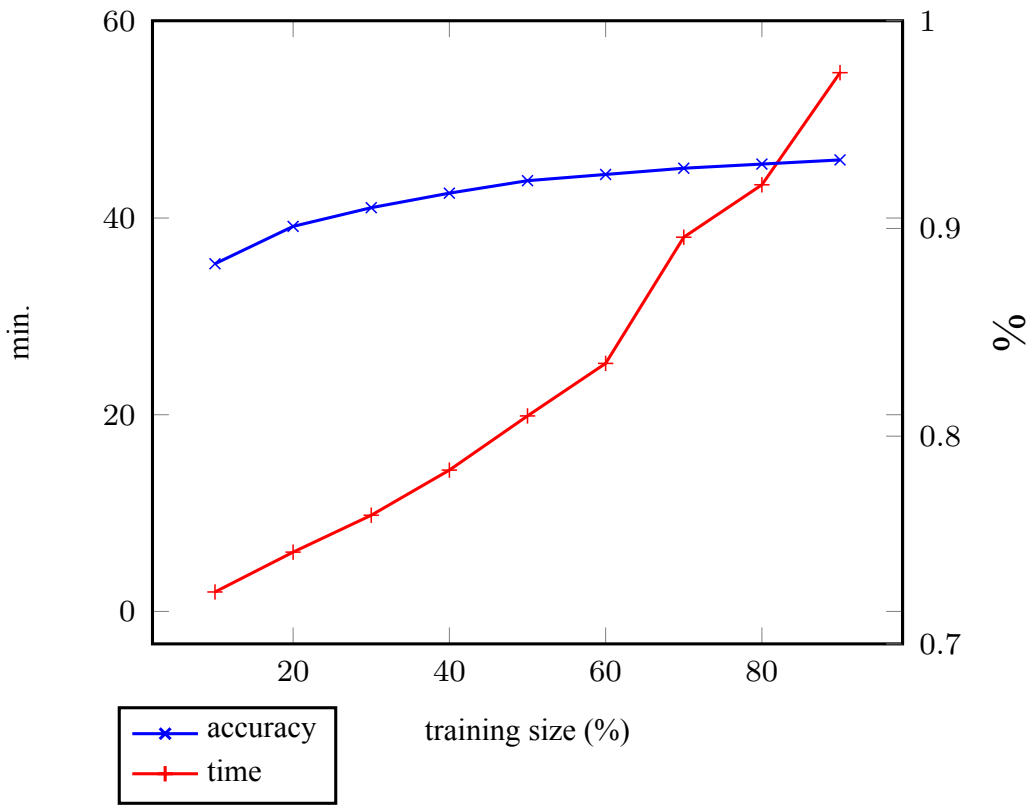


Figure – Accuracy and time of training

Experimental results

Error Type	Ratio
Tone Only	**58.52%**
Position Only	1.17%
Tone and Position	0.023%
Silence	**40.08%**

Table – Error by type (insertion operation)

		Predicted			
Actual		◌́	◌̀	◌̂	◌̌
	◌́	0.9541	**0.0438**	0.0021	0.0000
	◌̀	**0.0841**	0.9141	0.0015	0.0003
	◌̂	0.0035	**0.0322**	0.9643	0.0000
	◌̌	0.0000	**0.0952**	0.0000	0.9048

Table – Confusion matrix on prediction of tone markers

Plan

Projet LANGAS

- *Lenguas generales de América del Sur*
  `http://www.langas.cnrs.fr`
  - Projet SHS, langues **générales** suite à la colonisation
  - Langues : **quechua**, aymara, tupi, **guarani**

Projet LANGAS

- *Lenguas generales de América del Sur*
  `http://www.langas.cnrs.fr`
  - Projet SHS, langues **générales** suite à la colonisation
  - Langues : **quechua**, aymara, tupi, **guarani**
- Besoins SHS
  - Numérisation d'écrits
  - **Stockage / encodage des corpus**
  - **Indexation et recherche dans les corpus**
  - Vérification de la cohérence
  - Analyses outillées

$\Rightarrow$ Sollicitation de TAL (Johanna Cordova)

Aperçu des données

Lang.	Texts	Toks pal.	Toks translit.
Guarani	80	29,583	35,035
Quechua	31	250,593	113,547
Tupi	6	2700	NA

Table – Composition du corpus

Diacritic	Diacritizations
acute	á é í ó ú ý
grave accent	è ì ò ù ỳ
circumflex	â ê î ô û ŷ
tilde	ã ẽ ĩ õ ũ ỹ
breve	ĕ ĭ ў
inverted breve	ȃ ȇ ȋ ȏ ȗ ȳ

Table – Encodage des diacritiques

⇒ Cohérence de l'encodage (norme UFT8)

⇒ Expansion de requêtes sur les diacritisations

Quechua : variantes dialectales

- Problématique
  - Nombreuses variantes

  $\Rightarrow$ Clustering de variantes
  - Langue agglutinante

  $\Rightarrow$ Racinisation par **analyse morphologique**
- Utilisation de clustering automatique

Before merging	570 clusters
After merging	500 clusters
Avg. clusters size	2.16
Avg. inner distance	1.23
Most freq. equival.	ŝ $\sim$ s (164 cl.) ch $\sim$ ĉ (71 cl.) ch' $\sim$ ch $\sim$ chh (67 cl.) k' $\sim$ kh (53 cl.) p' $\sim$ p $\sim$ pp $\sim$ ph (56 cl.)

$\Rightarrow$ Expansion de requêtes sur les variantes

Expansion de requêtes : résultats

Word	Without exp.	With exp.	Ratio	Var
Quechua				
ñuqa	321	333	3.6%	4
kawsa	296	299	1%	2
yacha	230	303	33%	2
sunqu	158	415	8.8%	2
simi	110	371	70.3%	2
mikuy	17	66	74.2%	2
llamka	4	58	93.1%	3
ñisqa	36	485	96.5%	4
Guarani				
tupa	2	321	99,4%	2
nande	2	311	96,2%	2
guasu	135	156	6,5%	2
teko	138	139	0,3%	2
rera	11	57	14,3%	2

Translittération : paléographie vs moderne

- Version des textes
  - Version paléographique : transcription exacte

  $\Rightarrow$ Potention besoin d'**OCR**

  $\Rightarrow$ Peu lisible / uniforme / compréhensible

Translittération : paléographie vs moderne

- Version des textes
  - Version paléographique : transcription exacte

  $\Rightarrow$ Potention besoin d'**OCR**

  $\Rightarrow$ Peu lisible / uniforme / compréhensible
  - Version moderne ou **translittérée**
  - Entre translittération (écriture) et traduction (langue)

  $\Rightarrow$ Coûteux, vérifications nécessaires
- Exemple

Palaeography	Transliteration
Aiporano condequatia oremoñe mombĭaete, ore mboyeçaereco ete rano, y yabaiete orebe yacaho haguã,	Aipo rano ko nde kuatia oremoñemomby'a ete, ore mbojesaereko ete rano, ijavai ete oréve jakaho haguã,

Plan

Contexte

- Collaboration historique P8 / ERTIM / TSU (Tbilissi, AUF)

$\Rightarrow$ Enseignement de l'informatique

- 2019 : quelles avancées TAL pour le géorgien ?
  - Encodage : semble ok (UTF8)
  - Corpus : *Georgian National Corpus* `http://gnc.gov.ge/gnc`
  - Outils : bien peu, sauf standards / multilingues (Tesseract)

Objectifs du travail préliminaire

- Reconnaissance d'entités nommées
  - Essentiellement : les **noms propres**
  - Outil disponible dans de nombreuses langues
  - Complexité limitée (quoique)
  - Utilité incontestable (prétraitement)
- Prérequis
  - Lexiques
  - Corpus annotés
  - Outil symbolique / statistique
  - (Plongements de mots)

Projet étudiant

- Logiciel open source
  - Architecture en place (apprentissage)
  - Extraction automatique d'entités (MediaWiki)
  - Quelques textes annotés (BRAT)

⇒ `https://github.com/eldams/Georgian-NERD`

⇒ Annotation automatique... mais surtout des erreurs :(!

⇒ Problème : ambiguïtés, features, volumes de données

Suites du projet

- Utilisation de l'outil existant
  - Meilleure extraction des lexiques
  - Annotation de données en volume
- Amélioration de l'outil
  - Exploitation des plongements
  - Descriptions linguistiques
  - Autres méthodes / sources pour collecter les entités

Plan

Conclusions

- Ressources basiques pour traiter une langue
  - Un encodage fiable, consistant, cohérent (UTF8)

  $\Rightarrow$ Aussi en diachronie (manuscrits anciens)
  - De **grands** volumes de données (corpus)
  - Outils : segmentation, lemmatisation, morphologie, entités
- Applications qui peuvent être construites
  - Translittération
  - Moteurs de recherche
  - Reconnaissance de caractères
  - Reconnaissance vocale
  - Traduction / résumé

$\Rightarrow$ Jamais 100% : connaître et réduire les taux d'erreurs