Indexation automatique du patrimoine oral africain

A. NIMAAN [1,2] P. NOCERA [1] J-F. BONASTRE [1]

[1] Laboratoire d'Informatique d'Avignon (UAPV) —
[2] Institut des Sciences et des Nouvelles Technologies (CERD)

21 Juin 2007 - Grenoble

Sommaire

Sommaire

Introduction
Langue somalienne
Reconnaissance du somalien
Constitution de corpus
Modélisation acoustique
Reconnaissance de la parole lue
Archives du patrimoine oral
Décodage en racines
Décodage hybride
Recherche d'information
Choix des requêtes
Recherche en mots
Recherche en racines
Recherche hybride
Requêtes QOV
Conclusions

Outline

Introduction

- patrimoine culturel, historique et scientifique des pays africains ;
- menace de disparition ;
- organisations nationales et internationales (UNESCO, Union Africaine, etc.) ;
- vastes programmes de vulgarisation du patrimoine oral ;
- importantes archives audio disponibles.

Introduction

- programme de numérisation ;
- meilleure accessibilité ;

Introduction

- programme de numérisation ;
- meilleure accessibilité ;

$\Rightarrow$ numérisation = problème d'ordre logistique ;

Introduction

- programme de numérisation ;
- meilleure accessibilité ;

$\Rightarrow$ numérisation = problème d'ordre logistique ;

$\Rightarrow$ exploitation de bases de données audio de grandes tailles = outils informatiques de haut niveau ;

Introduction

- programme de numérisation ;
- meilleure accessibilité ;

$\Rightarrow$ numérisation = problème d'ordre logistique ;

$\Rightarrow$ exploitation de bases de données audio de grandes tailles = outils informatiques de haut niveau ;

$\Rightarrow$ outils de transcription et d'indexation automatiques.

Introduction

Problématiques :

- système complet de Reconnaissance Automatique de la Parole (RAP) pour une nouvelle langue ;
- recherches d'information à partir des transcriptions automatiques ?

$\Rightarrow$ manque ou l'insuffisance de corpus (peu ou pas d'outils informatiques) ;

$\Rightarrow$ performances ? (patrimoine oral) ;

$\Rightarrow$ quelles stratégies ?

Langue somalienne

- Djibouti, Somalie, Ethiopie et le Kenya ;
- 11 et 13 millions (Ethonologue 2005) ;
- afro-asiatique, couchitique, couchitique-Est ;
- (Saeed, 1993) et (Andrzejewski et Lewis, 1964) ;
- alphabet latin = écriture officielle (1972) ;
- 22 consonnes (', B, T, J, X, KH, D, R, S, SH, DH, C, G, F, Q, K, L, M, N, W, H, Y) ;
- 10 voyelles (A, E, I, O et U) + (AA, EE, II, OO et UU) ;
- structures syllabiques (CVC, CV, VC, V) nommées "racines" ;
- organe de régulation ou de standardisation de l'orthographe.

Outline

Constitution de corpus

- 2 corpus acoustiques (*ASAAS* et *RTD*)
- 1 corpus textuel (*WARGEYS*)

Constitution de corpus

- 2 corpus acoustiques (*ASAAS* et *RTD*)
- 1 corpus textuel (*WARGEYS*)

⇒ **ASAAS** : corpus de parole lue (articles de journaux 2002-2004) d'environ ≈ 10 heures, manuellement transcrit.

⇒ deux parties : apprentissage (8 heures et 57 minutes) et developpement (1 heure et 29 minutes) ;

Constitution de corpus

- 2 corpus acoustiques (*ASAAS* et *RTD*)
- 1 corpus textuel (*WARGEYS*)

⇒ **ASAAS** : corpus de parole lue (articles de journaux 2002-2004) d'environ $\approx$ 10 heures, manuellement transcrit.

⇒ deux parties : apprentissage (8 heures et 57 minutes) et developpement (1 heure et 29 minutes) ;

⇒ **RTD** = corpus d'émissions culturelles du patrimoine djiboutien d'environ $\approx$ 1 heure, manuellement transcrit.

Constitution de corpus

- 2 corpus acoustiques (*ASAAS* et *RTD*)
- 1 corpus textuel (*WARGEYS*)

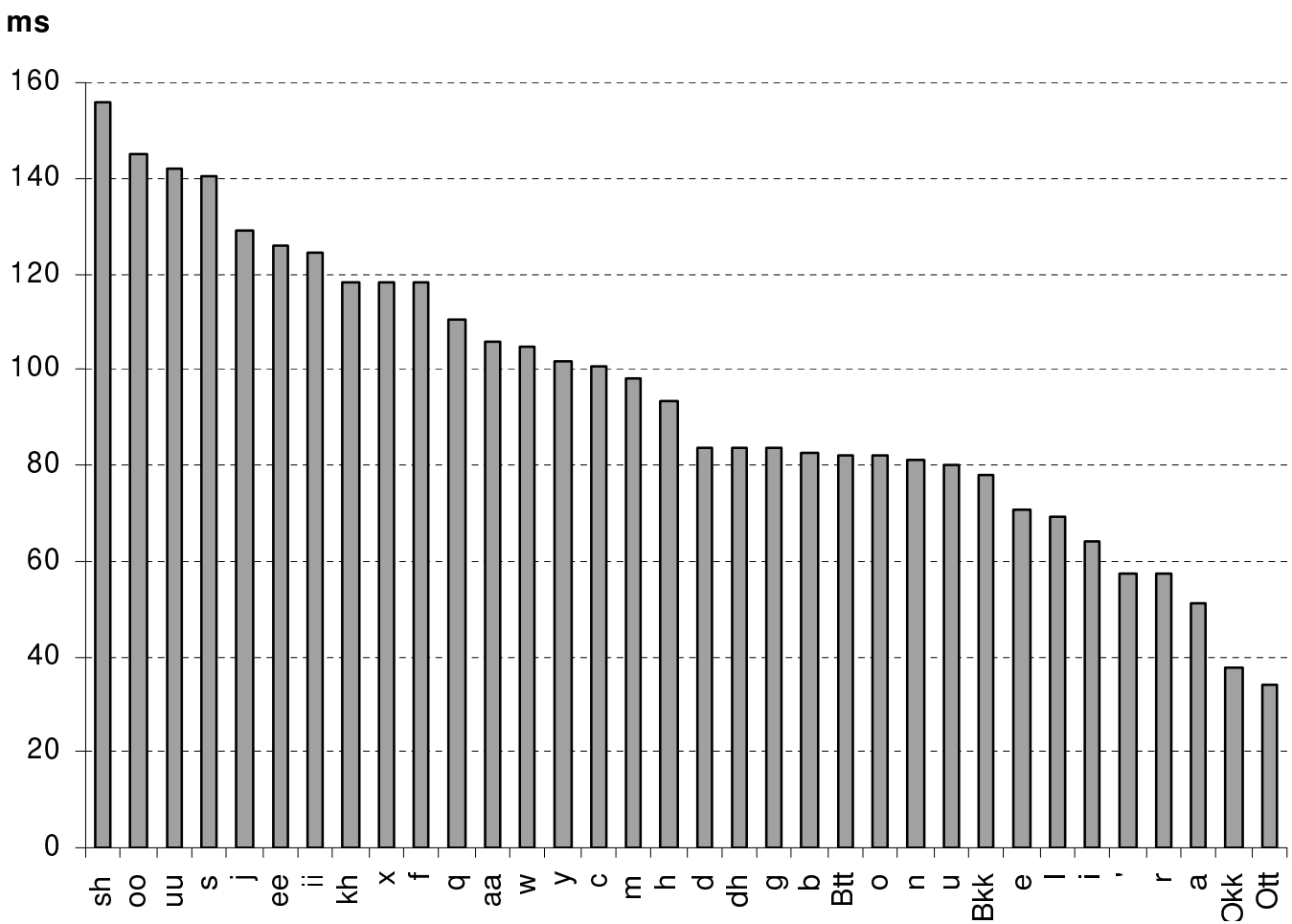
⇒ **ASAAS** : corpus de parole lue (articles de journaux 2002-2004) d'environ $\approx$ 10 heures, manuellement transcrit.

⇒ deux parties : apprentissage (8 heures et 57 minutes) et developpement (1 heure et 29 minutes) ;

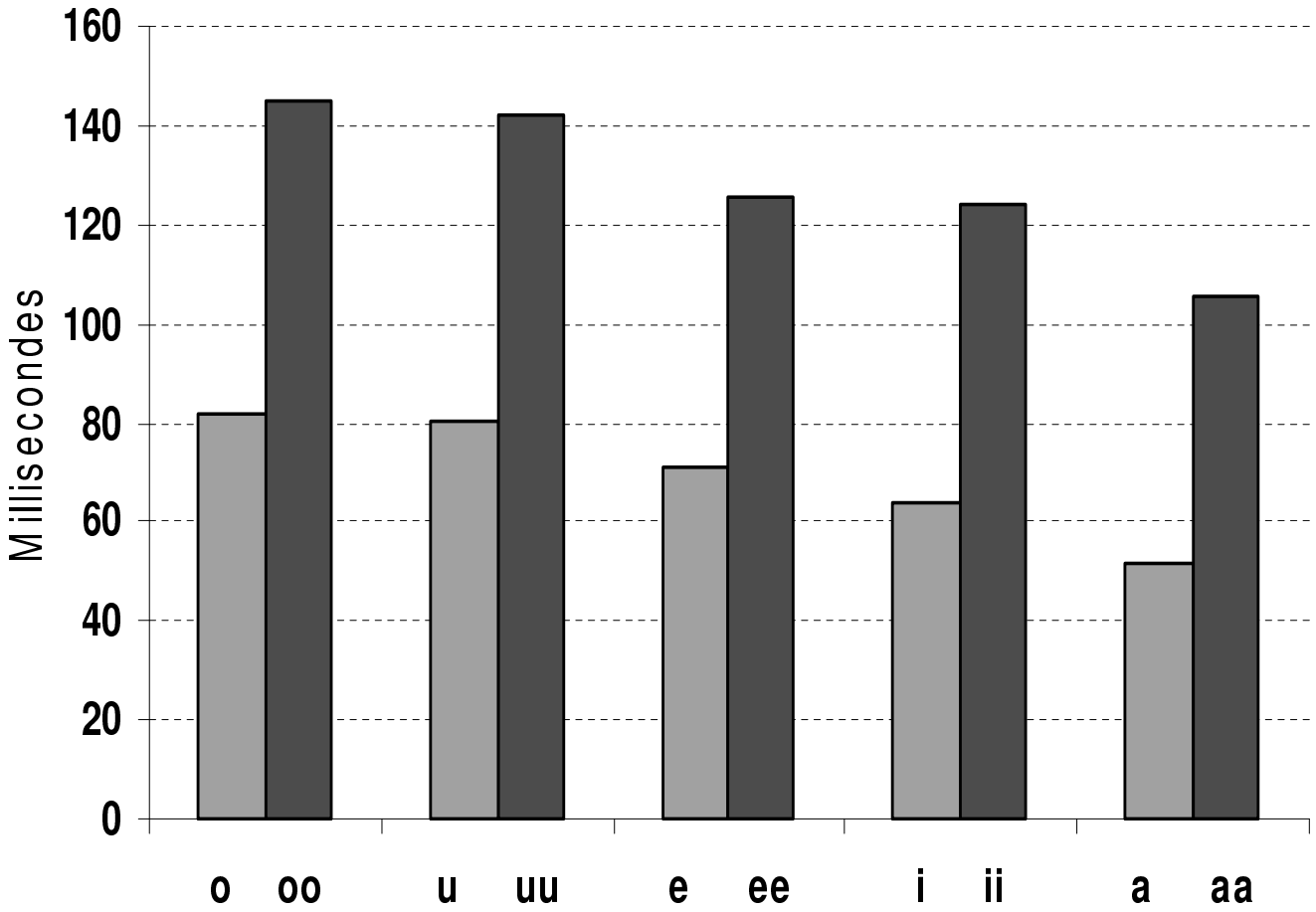
⇒ **RTD** = corpus d'émissions culturelles du patrimoine djiboutien d'environ $\approx$ 1 heure, manuellement transcrit.

⇒ **WARGEYS** = corpus textuel de type journalistique (2002-2004) recoltés sur Internet.

Caractéristiques du corpus ASAAS



Caractéristiques du corpus ASAAS



Modélisation acoustique

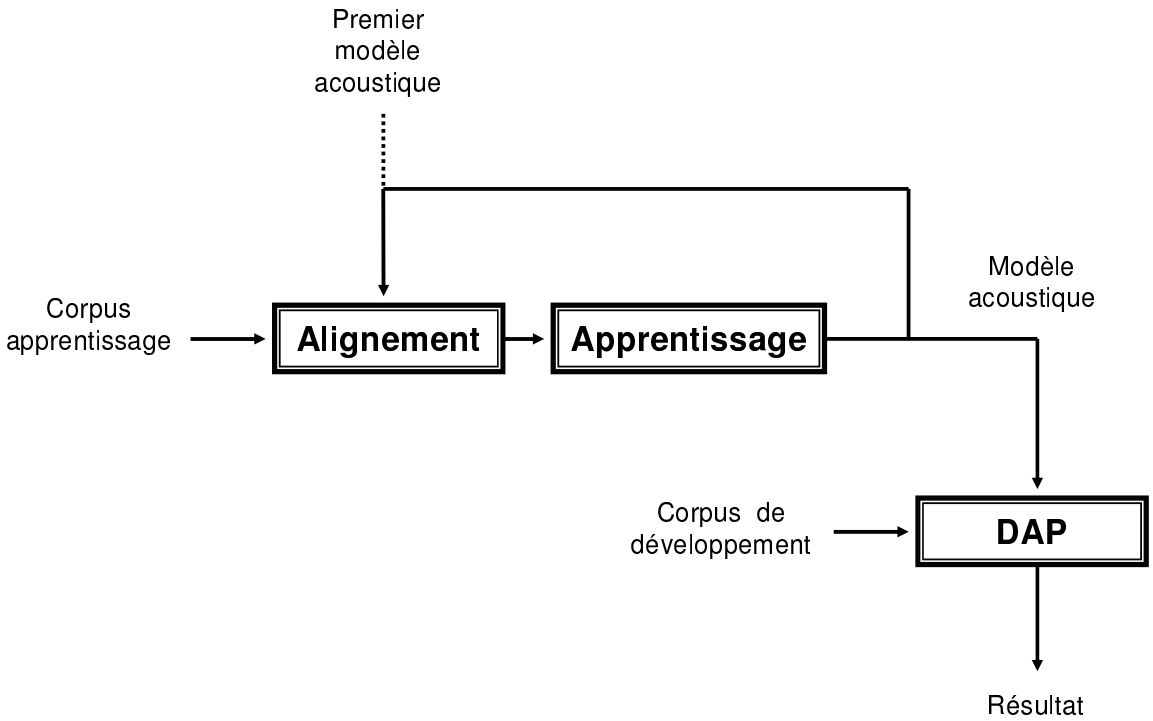
Deux méthodes de mise en relation des phonèmes de la langue somalienne et ceux du français ont été comparées :

- "**experte**" : basée sur des connaissances *a priori*.
- "**automatique**" : basée une matrice de confusion.

Modélisation acoustique

Deux méthodes de mise en relation des phonèmes de la langue somalienne et ceux du français ont été comparées :

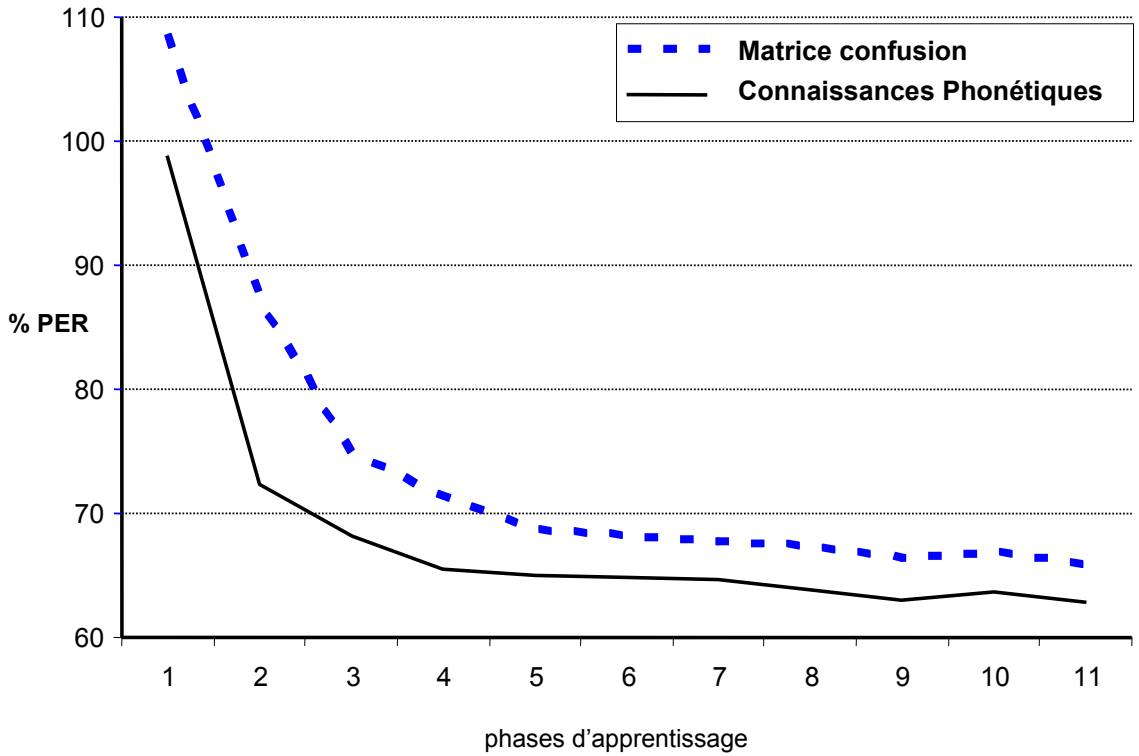
- "**experte**" : basée sur des connaissances *a priori*.
- "**automatique**" : basée une matrice de confusion.



Modélisation acoustique

Comparaison des deux méthodes : Experte et automatique.

PER : Phoneme Error Rate



Reconnaissance de la parole lue

Premières expériences de RAP :

- corpus de développement ASAAS-Dev ;
- LM_{20k}, OOV = 4,90% (Out Of Vocabulary) ;
- Décodeur Speeral du LIA.

Reconnaissance de la parole lue

Premières expériences de RAP :

- corpus de développement ASAAS-Dev ;
- LM_{20k}, OOV = 4,90% (Out Of Vocabulary) ;
- Décodeur Speeral du LIA.

Cor (%)	Sub (%)	Supp (%)	Ins (%)	WER (%)
76,4	20,8	2,8	4,7	28,3

Reconnaissance de la parole lue

Premières expériences de RAP :

- corpus de développement ASAAS-Dev ;
- LM_{20k}, OOV = 4,90% (Out Of Vocabulary) ;
- Décodeur Speeral du LIA.

Ref : **GUDDOOMIYAHA** gobolka oo uu **WERIYAHAYAGU** wax
Hyp : **GUDOOMIYAHA** gobolka oo uu **WARIYAHAYAGU** wax

Ref : ka **WEYDIIYAY** ARIMAHA AY ka wada hadleen WUXUU
Hyp : ka **WAYDIIYAY** MASTAR ILAAHAY ka wada hadleen UGU

Ref : sheegay in waqti kale ay ** ***** U **BALLAMEEN**
Hyp : sheegay in waqti kale ay KU TIMID **BALAMEEN**

Ref : **DHAMMAYSTIRKA** HESHIISYO ***** hore U
Hyp : **DHAMAYSTIRKA** BISHII SIIYO hore UGU

Ref : dhexmaray oo * aanu **FAAH** **FAAHIN**
Hyp : dhexmaray oo U aanu **** **FAAHFAAHIN**

Normalisation

Le manque de normalisation (standardisation) de l'orthographe :

- problème non spécifique à la langue somalienne ;
- plupart des langues africaines écrites depuis peu ;
- manque de normalisation $\neq$ fautes d'orthographes ;
- plusieurs transcriptions d'un même mot ;
- problèmes des noms propres (*jabuuti*, *jabbuuti*, *jabuutti*, *jibuuti*, etc.)
- mots empruntés et "africanisés" (*biirootabaa*, *biirootamaa*, *biiroo tabaa*, etc. [bureau tabac])

Normalisation

Règles de normalisation (langue somalienne) :

- transcription la plus fréquente (WARGEYS) $\Rightarrow$ transcription retenue (n'est pas forcément la meilleure) ;
- mots composés (*faah faahin*, *faahfaahin*) $\Rightarrow$ mots composés retenues (*faah faahin*) - formes fléchies - ;
- lettres doubles (très fréquentes) $\Rightarrow$ lettres simples ;

Reconnaissance de la parole lue

Normalisation des sorties du système et des références.

- WER = 21,5% ;
- gain relatif de 24%.

Corpus normalisé	Cor	Sub	Supp	Ins	WER
AUCUN	76,4	20,8	2,8	4,7	28,3
ASAAS-Dev	83,7	14,7	1,6	5,2	21,5

Reconnaissance de la parole lue

Normalisation du corpus WARGEYS.

- WER = 20,2% ;
- gain relatif de 28%.

Corpus normalisés	Cor	Sub	Supp	Ins	WER
AUCUN	76,4	20,8	2,8	4,7	28,3
ASAAS-Dev	83,7	14,7	1,6	5,2	21,5
WARGEYS, ASAAS-Dev	85,1	13,4	1,5	5,3	20,2

Conclusions parole lue

Le système de reconnaissance de la langue somalienne :

- difficultés liées au manque de corpus ;
- problèmes de modélisations ;
- Normalisation de l'orthographe ;
- Taux d'erreur satisfaisant.

Conclusions parole lue

Le système de reconnaissance de la langue somalienne :

- difficultés liées au manque de corpus ;
- problèmes de modélisations ;
- Normalisation de l'orthographe ;
- Taux d'erreur satisfaisant.

Réglage des paramètres :

- modèles acoustiques "robustes"
- 128 gaussiennes par état ;
- corpus textuels sont normalisés (orthographe) ;
- modèle de langage trigrammes.

Reconnaissance du patrimoine oral

Différents thèmes du Corpus RTD :

Sujet	**Traduction**	
Bilawgii fanka	Les chants traditionnels	*XIX^e*
Munafaacadka badda	La mer	*XX^e*
Taariikhda geeska afriika	Histoire de la corne de l'Afrique	
Aale Boore	*Aale Boore*	*XVI^e*
Xariirta	La soie	
Nabiga	Le prophète	*VII^e*
Cabdiraxmaan saylici	*Cabdiraxmaan saylici*	*XIX^e*
Furitaanka	Le divorce	*XX^e*

Reconnaissance du patrimoine oral

Reconnaissance du patrimoine oral :

- corpus RTD ;
- LM_{20k} ;
- Décodeur Speeral du LIA.

Reconnaissance du patrimoine oral

Reconnaissance du patrimoine oral :

- corpus RTD ;
- LM_{20k} ;
- Décodeur Speeral du LIA.

Cor (%)	Sub (%)	Supp (%)	Ins (%)	WER (%)
46,6	46,4	7,0	8,7	62,1

Reconnaissance du patrimoine oral

Reconnaissance du patrimoine oral :

- corpus RTD ;
- LM_{20k} ;
- Décodeur Speeral du LIA.

Cor (%)	Sub (%)	Supp (%)	Ins (%)	WER (%)
46,6	46,4	7,0	8,7	62,1

- résultat était prévisible ;
- impossibilité de trouver des corpus d'entrainement adaptés ;
- OOV = 12,48% ; (noms de lieux, personnages, évènements, etc).

Reconnaissance du patrimoine oral

Reconnaissance du patrimoine oral :

- corpus RTD ;
- LM_{20k} ;
- Décodeur Speeral du LIA.

Cor (%)	Sub (%)	Supp (%)	Ins (%)	WER (%)
46,6	46,4	7,0	8,7	62,1

- résultat était prévisible ;
- impossibilité de trouver des corpus d'entrainement adaptés ;
- OOV = 12,48% ; (noms de lieux, personnages, évènements, etc).

Représentation suffisamment robuste aux décalages temporels et thématiques.

Décodage en racines

Reconnaissance en syllabes (”racines”) = voie intéressante.

- langue somalienne = langue agglutinante ;
- mots composés de suites de ”racines-sylabbes” (CVC, CV, VC, V)
- nombre de racines limité ($\approx$ 5 000) ;
- racines sont à la base de la formation des mots (anciens ou nouveaux) ;
- représentation en racines illisible $\longrightarrow$ unités d'indexation.

Décodage en racines

- lexique de reconnaissance (4 400 racines) ;
- modèle de langage trigrammes de racines (189 000 bigrammes, 996 000 trigrammes) ;

Décodage en racines

- lexique de reconnaissance (4 400 racines) ;
- modèle de langage trigrammes de racines (189 000 bigrammes, 996 000 trigrammes) ;
- taux de racines OOV presque nulle (0,03%) ;
- perplexité non normalisée = 19,05 (corpus RTD).

Décodage en racines

Résultat du décodage en racines du corpus RTD (RER = Root Error Rate).

Cor (%)	Sub (%)	Supp (%)	Ins (%)	RER (%)
57,2	32,3	10,5	4,2	47,0

Décodage en racines

Résultat du décodage en racines du corpus RTD (RER = Root Error Rate).

Cor (%)	Sub (%)	Supp (%)	Ins (%)	RER (%)
57,2	32,3	10,5	4,2	47,0

Emission *SOYAAL "Cabdiraxmaan Saylici"*

REF: ** een ** mag aal AD a har AR oo ah ay meel xar uun AH OO wayn oo lag u bar O diin ta say lac OO meel BAAY AC **MUSH TAR I** AH ay **TAJ OOR a** ber ber a dhul kaas oo dhan il aa mus aw AC **dhul kaas O** OO DHAN oo WAD A ah AA dhul kii MA GAR AN AY saa ** diin TA IY O BAAY AC **MUSH TAR ig *** a** GEES KA AF RIIK A ay AY ay IM AN ay EEN ay ** AAB a gid ig ood WAY WAD A ON OOB een

HYP: EE een IN mag aal AH a har ** oo ah ay meel xar uun ** ** wayn oo lag u bar TO diin ta say lac A meel **** ** **** AAN U SHEEG ay KA JIR a ber ber a dhul kaas oo dhan il aa mus aw A dhul kaas * ** **** oo LA MOOD ah A dhul kii ** MAQ AL NAY saa EE diin ** ** * DAY DA CUS UB ig IIS a **** ** SID AA YEEL ay GAY ay ** SIIN ay *** ay AA BAA a gid ig ood A WAW AL O NOO een

Emission *WAR IYO WAAYO ARAG "Furitaanka"*

REF: ar in la xidh iidh a **MUJ tam ac een a** wax AN a weey aan fur IT AAN kii **** OO BAT AY ee **od ay AD II** war iy o waay o ar ag wax ay dar eem een in im IK A rag u ay **mas uul iy ad ood iib A DAY AC EEN** oo DUM AR kii sid ii loog u sam ri jir ay iy o sid ii reer ah a loo wad I jir ay ba ay ka tag een oo wix ii BA u noq ** DAY is laan tii WAAN fur ay iy o WAAN FUR ay aa

HYP: ar in la xidh iidh a MUSH tam ac een a wax AAN a weey aan fur ** TAN kii KOOB AN TAY EE ee od ay ** GII war iy o waay o ar ag wax ay dar eem een in im O KA rag u ay mas uul iy ad

Décodage en racines

Résultat du décodage en racines du corpus RTD (RER = Root Error Rate).

Cor (%)	Sub (%)	Supp (%)	Ins (%)	RER (%)
57,2	32,3	10,5	4,2	47,0

Exemples de mots OOV reconnus par le décodage en racines.

mot OOV	LM_{20k}	RLM
asnaamtaasi	wasaaradaasi	as naam ta si
tafaraaruqa	taf abaabulka	taf ar aar uq a
faaqidi	nafaqada	aq ad i
(shiinaha) qudhooda	bishii lagu looga	bish iib a qudh ood a
(laba) dakhare	labadaba	lab ad a sar e

Décodage en racines

Comparaison entre le décodage en mots et celui en racines.

- hypothèses mots décomposées en racines ;
- fichiers références décomposées en racines ;
- taux d'erreur obtenu (WRER = Word-Root Error Rate).

Décodage en racines

Comparaison entre le décodage en mots et celui en racines.

- hypothèses mots décomposées en racines ;
- fichiers références décomposées en racines ;
- taux d'erreur obtenu (WRER = Word-Root Error Rate).

Racines (RER)	47,0
Mots-décomposés (WRER)	46.4

Décodage en racines

Comparaison entre le décodage en mots et celui en racines.

- hypothèses mots décomposées en racines ;
- fichiers références décomposées en racines ;
- taux d'erreur obtenu (WRER = Word-Root Error Rate).

Racines (RER)	47,0
Mots-décomposés (WRER)	46.4

WRER légèrement plus faible que le RER, malgrè le taux de OOV $\Rightarrow$ profondeur du modèle de langage en mots.

Décodage hybride

Idée : bénéficier des avantages du modèle de langage mots (WLM) et de la gestion des OOV du modèle de langage en racines (RLM)

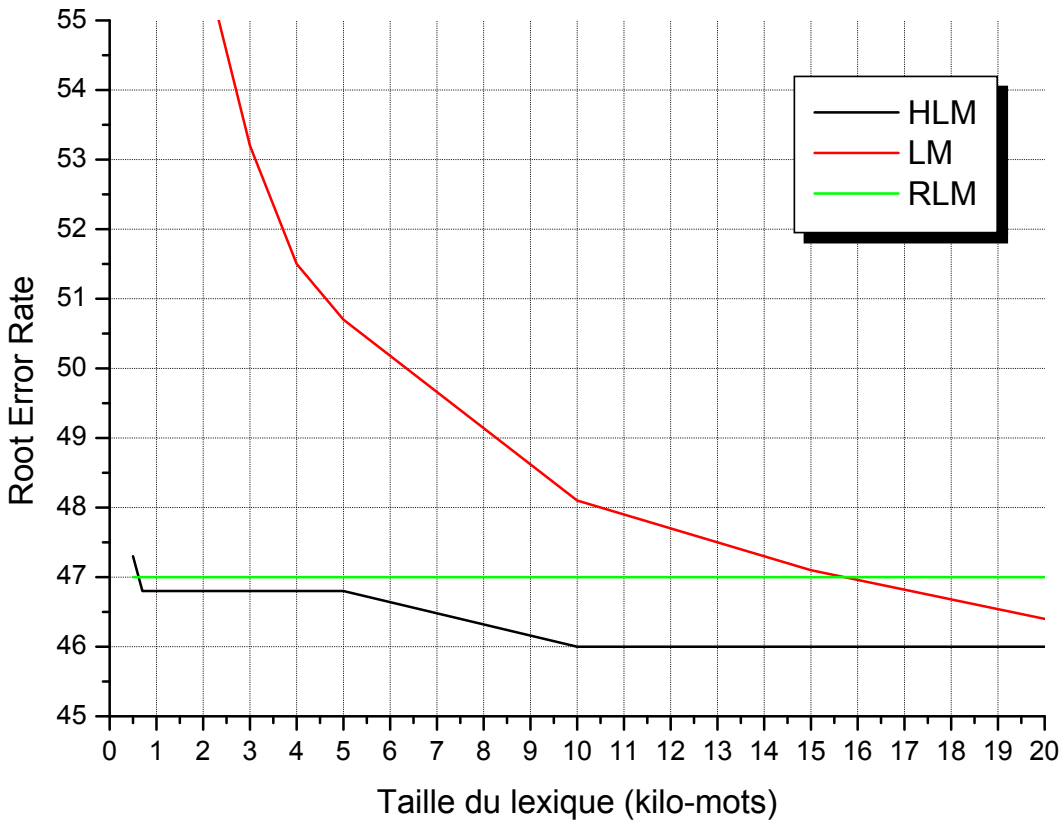
- modèle de langage hybride (*HLM*) ;
- lexique restreint n mots et le reste en racines HLM_n ;
- n variant de 200 à 20 000 mots ;
- lexique de reconnaissance de taille $\approx n + 5k$.

Décodage hybride

- RER = 46% pour HLM_{20k} ;
- systèmes hybrides sont meilleurs que ceux en mots ou en racines quelque soit la taille du lexique.

Décodage hybride

- RER = 46% pour HLM_{20k} ;
- systèmes hybrides sont meilleurs que ceux en mots ou en racines quelque soit la taille du lexique.



Décodage hybride

- RER = 46% pour HLM_{20k} ;
- systèmes hybrides sont meilleurs que ceux en mots ou en racines quelque soit la taille du lexique.

Certains mots OOV reconnus par les systèmes basés sur les HLM_n.

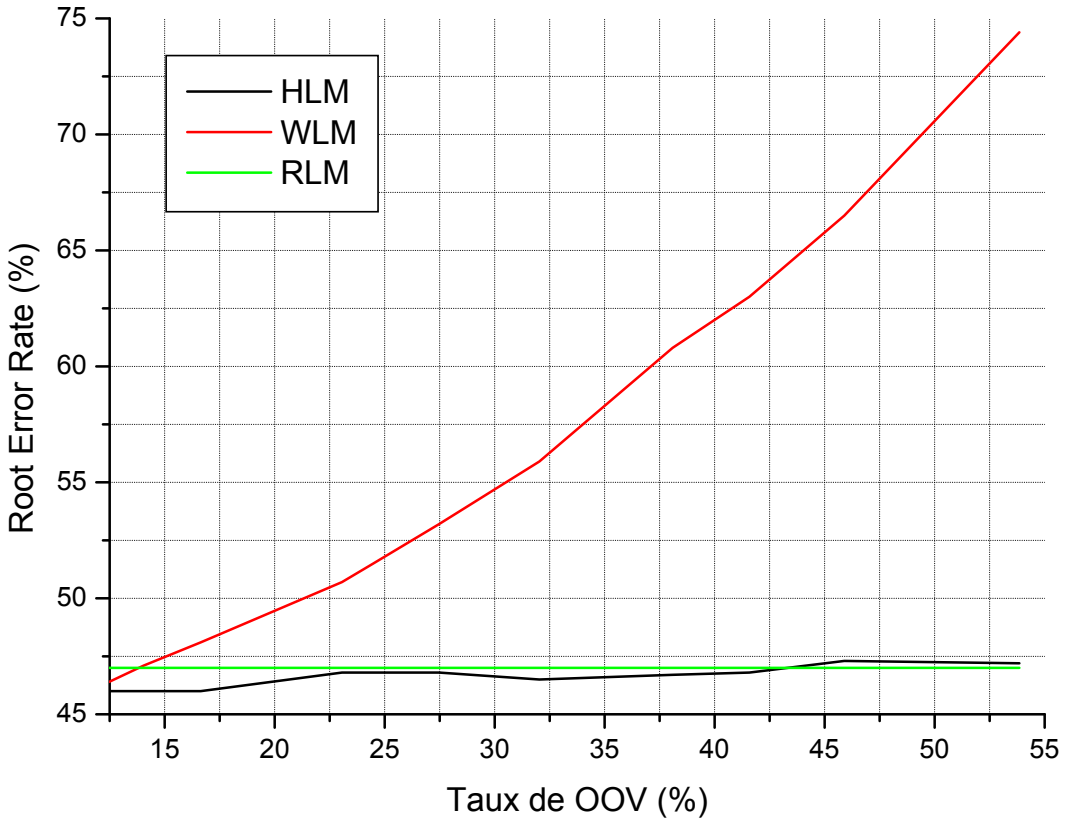
Reference	WLM 20k	HLM 20k	HLM 1k	RLM
asnaamtaasi	wasaaradaasi	as naam ta si	as naam ta si	as naam ta si
tafaraaruqa	taf abaabulka	taf ar aar uq a	taf ar aar uq a	taf ar aar uq a
faaqidi	nafaqada	faaq id i	faaq id i	aq ad i
(shiinaha) qudhooda	bishii lagu looga	shiinaha qudh ood a	bishii ina qudh ood a	bish iib a qudh ood a
(laba) dakhare	labadaba	laba dakh ar e	labada kale	lab ad a sar e

Décodage hybride

- RER = 46% pour HLM_{20k} ;
- systèmes hybrides sont meilleurs que ceux en mots ou en racines quelque soit la taille du lexique.

Décodage hybride

- RER = 46% pour HLM_{20k} ;
- systèmes hybrides sont meilleurs que ceux en mots ou en racines quelque soit la taille du lexique.



Outline

Cadre expériemental

Corpus RTD.

Sujet	**Traduction**	
Bilawgii fanka	Les chants traditionnels	XIX^e
Munafaacadka badda	La mer	XX^e
Taariikhda geeska afriika	Histoire de la corne de l'Afrique	
Aale Boore	*Aale Boore*	XVI^e
Xariirta	La soie	
Nabiga	Le prophète	VII^e
Cabdiraxmaan saylici	*Cabdiraxmaan saylici*	XIX^e
Furitaanka	Le divorce	XX^e

Cadre expériemental

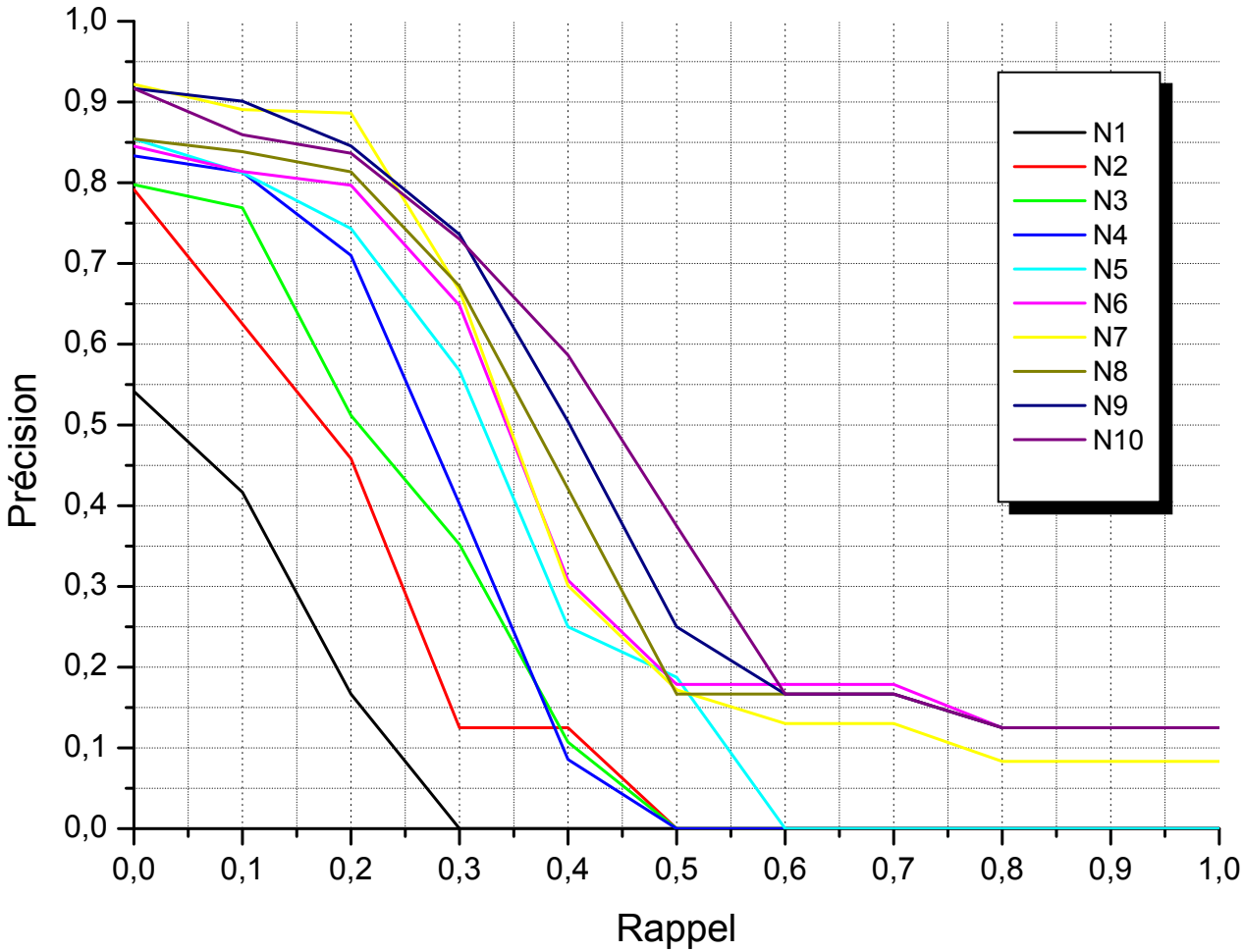
- requêtes presque idéales ;
- vecteur des 20 mots les plus pertinents (*tfidf*) ;
- moteur de recherche classique, basé sur la méthode vectorielle ;
- recherches en racines ou hybrides $\Rightarrow$ requêtes format adéquats ;

Choix des requêtes

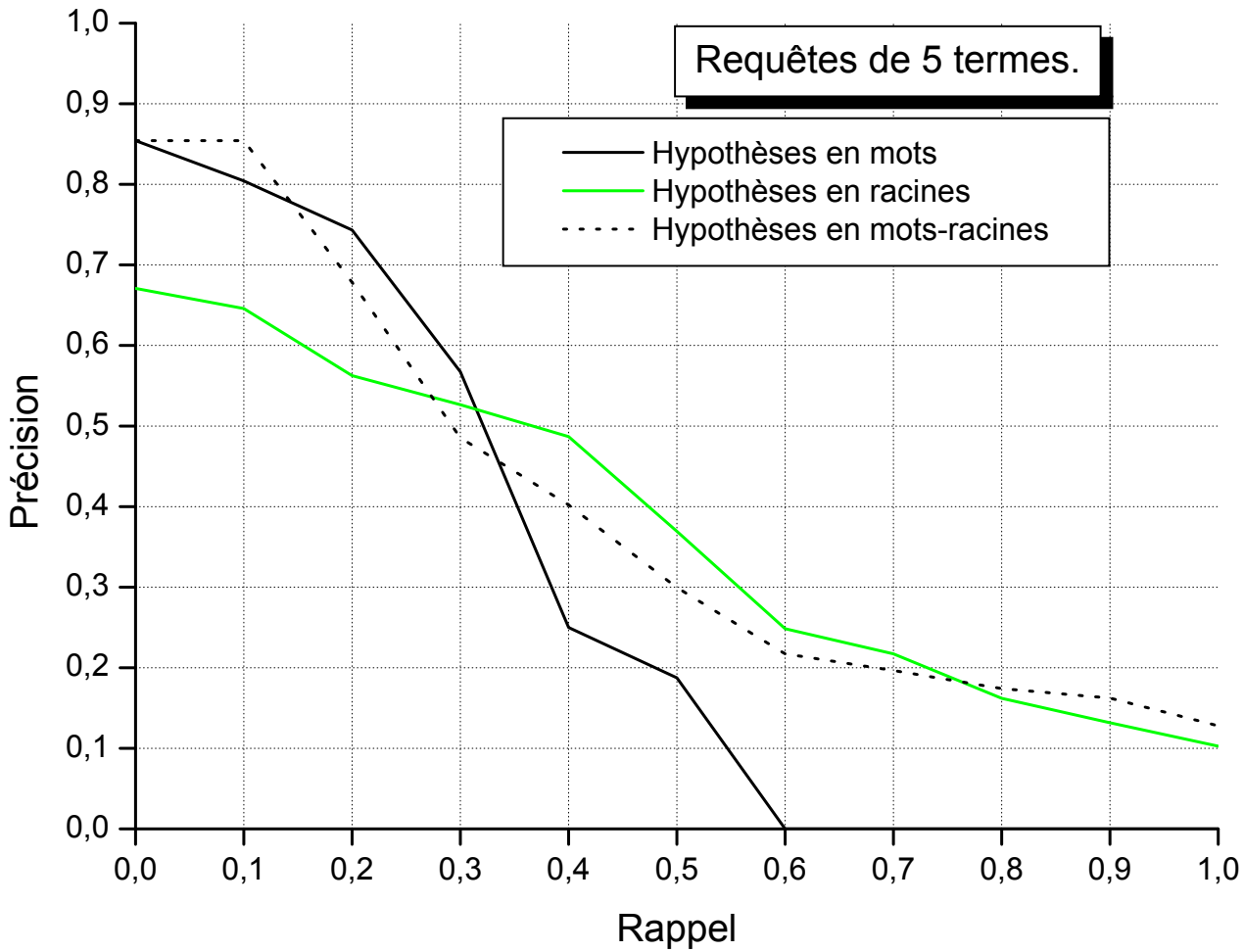
Nombre de QOV (Query Out-of-Vocabulary items).

	Requête	Nombre de termes	#QOV
1	Les chants traditionnels	20	9
2	Histoire de la corne de l'Afrique	20	2
3	La mer	20	4
4	La soie	20	4
5	*Aale Borre*	20	7
6	Le prophète	20	4
7	*Cabdiraxmaan saylici*	20	9
8	Le divorce aujourd'hui	20	3

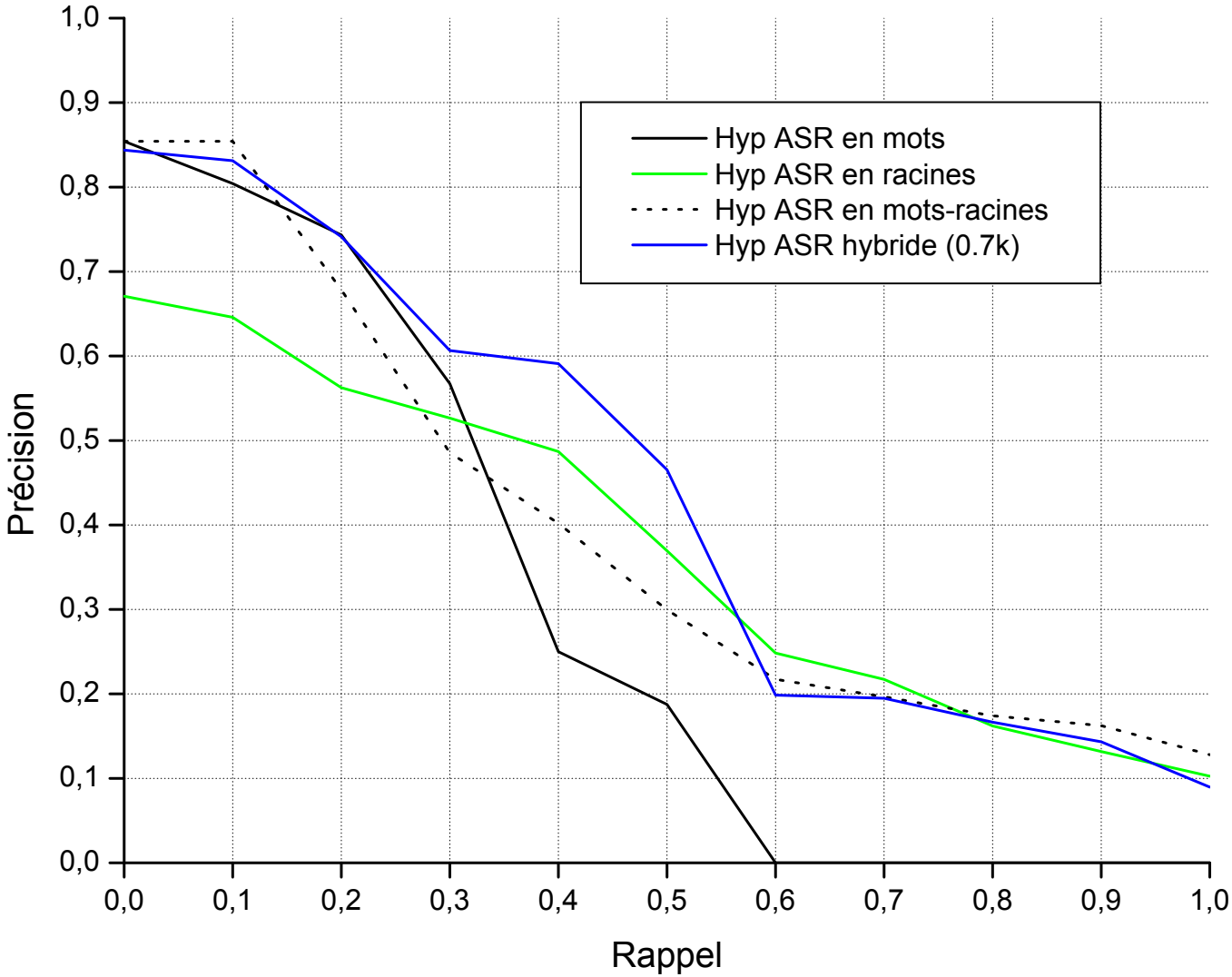
Recherche en mots



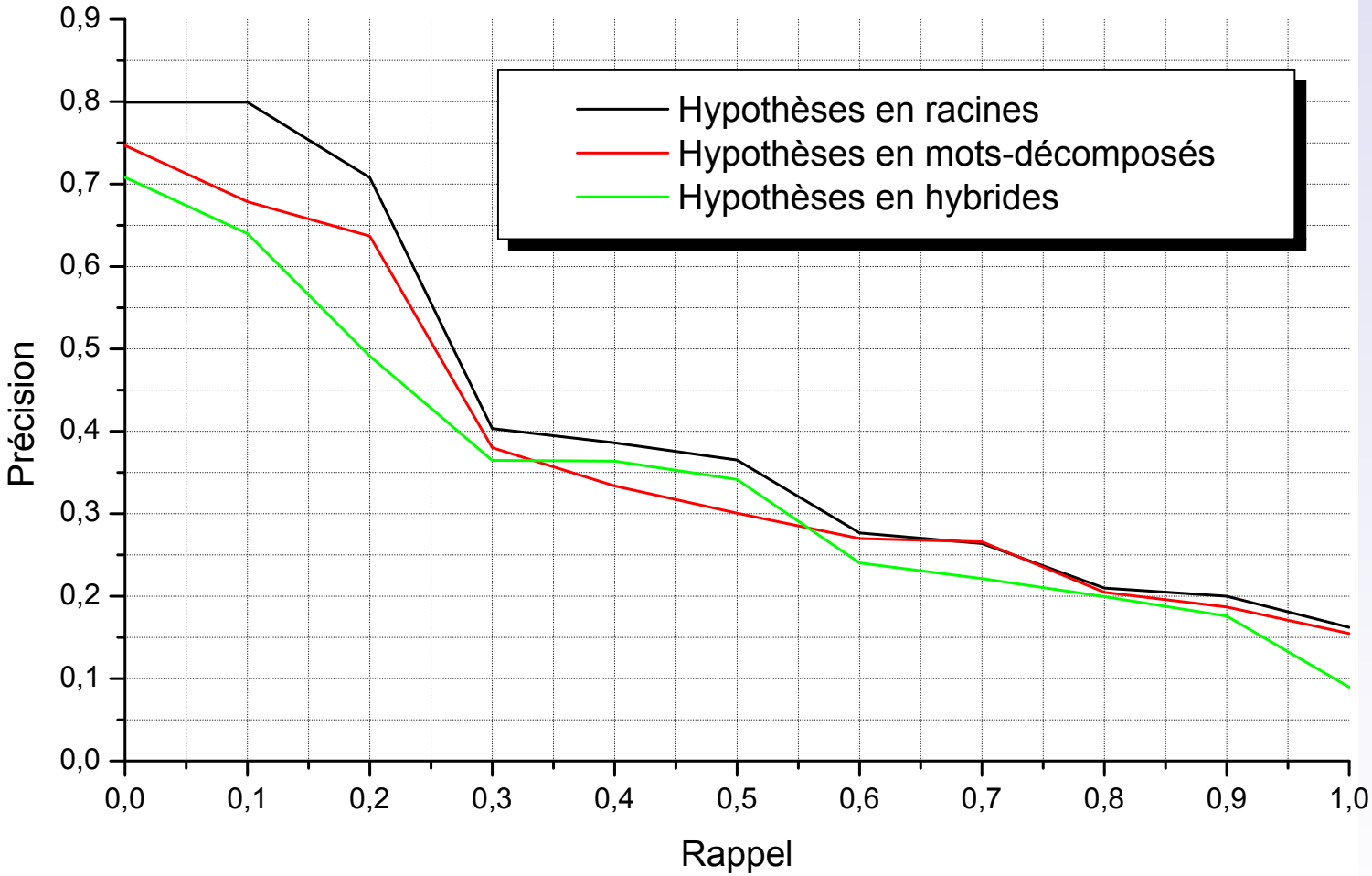
Recherche en racines



Recherche hybride



Requêtes QOV



Outline

Introduction

Reconnaissance du somalien

Recherche d'information

Conclusions

Conclusions

- corpus (Audio $\approx$ 10 heures) ;
- corpus textuel ($\approx$ 3 millions de mots) ;
- corpus RTD (extrait des émissions culturelles $\approx$ 1 heure) ;
- premier système de reconnaissance de la parole grand vocabulaire pour la langue somalienne ;
- Boîte à outils SOMTOOLS (phonétisation, décomposition en racines, transducteurs, etc.)
- méthode de création rapide d'un modèle acoustique pour une nouvelle langue ;
- difficultés liées au manque de normalisation de l'orthographe (28%) ;
- archives $\neq$ données apprentissages (distance thématique et temporelle) ;
- décodage en racines (hybrides) $\Rightarrow$ atténuer cette difficulté ;

Conclusions

- requêtes à fort taux de QOV (noms propres, événements historiques, etc.) ;
- recherche en mots ”limitée” par le couple OOV-QOV (petits rappels) ;
- recherche en racines (grands rappels) ;
- recherche ”hybride” (petits et grands rappels) ;
- R=50%, (requêtes=5mots, QOV=22%) $\Rightarrow$ P=18,75% (mots), P=36,94% (racines), P=46,53% (hybride 0.7k) ;
- R=50%, requêtes QOV $\Rightarrow$ P $\approx$ 30% (racines, hybrides) (P=0% pour la recherche en mots) ;
- résultats encourageants ;
- recherche ”hybride” = la mieux adaptée.