PRESERVATION OF AFRICAN CULTURAL HERITAGE BY AUTOMATIC TRANSCRIPTION OF AFRICAN LANGUAGES

NIMAAN A.[1], **NOCERA P.**[2]

1 Institut des Sciences et des Nouvelles Technologies de Djibouti
2 Laboratoire Informatique d'Avignon

UNIVERSITÉ D'AVIGNON
MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

LABORATOIRE D'INFORMATIQUE CERI
339 Chemin des Meinajariès
BP 1228
84911 AVIGNON CEDEX 09
Tél. + 33 (0)4 90 84 35 09
Fax. + 33 (0)4 90 84 35 01
secretariat-lia@univ-avignon.fr
http://www.lia.univ-avignon.fr

Laboratoire d'Informatique
Université d'Avignon

UNIVERSITÉ D'AVIGNON

MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Introduction

In Africa, the cultural, scientific and historical patrimonies are transmitted orally through generations.

This ancestral knowledge is disappearing

Most of the concerned countries, regional and international organizations (UNESCO) are elaborating policies to save this human richness.

LABORATOIRE D'INFORMATIQUE CERI
339 Chemin des Meinajariès
BP 1228
84911 AVIGNON CEDEX 09
Tél. + 33 (0)4 90 84 35 09
Fax. + 33 (0)4 90 84 35 01
secretariat-lia@univ-avignon.fr
http://www.lia.univ-avignon.fr

Laboratoire d'Informatique
Université d'Avignon

UNIVERSITÉ D'AVIGNON
MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Introduction

Two main issues:

- Saving : recording and digitalizing the oral patrimony
- Using : access the databases

LABORATOIRE D'INFORMATIQUE CERI
339 Chemin des Meinajariès
BP 1228
84911 AVIGNON CEDEX 09
Tél. + 33 (0)4 90 84 35 09
Fax. + 33 (0)4 90 84 35 01
secretariat-lia@univ-avignon.fr
http://www.lia.univ-avignon.fr
Laboratoire d'Informatique
Université d'Avignon

UNIVERSITÉ D'AVIGNON

MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Our objective

- Build a spoken document retrieval system for the Djibouti Republic.
  - Find the best speech representation for document retrieval

LABORATOIRE D'INFORMATIQUE CERI
339 Chemin des Meinajariès
BP 1228
84911 AVIGNON CEDEX 09
Tél. + 33 (0)4 90 84 35 09
Fax. + 33 (0)4 90 84 35 01
secretariat-lia@univ-avignon.fr
http://www.lia.univ-avignon.fr

Laboratoire d'Informatique
Université d'Avignon

UNIVERSITÉ D'AVIGNON
MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

LABORATOIRE D'INFORMATIQUE CERI
339 Chemin des Meinajariès
BP 1228
84911 AVIGNON CEDEX 09
Tél. + 33 (0)4 90 84 35 09
Fax. + 33 (0)4 90 84 35 01
secretariat-lia@univ-avignon.fr
http://www.lia.univ-avignon.fr

Laboratoire d'Informatique
Université d'Avignon

Outline

- The Somali language
- Corpus
- Speech Recognition System
- Experiments and improvements
- Conclusion

UNIVERSITÉ D'AVIGNON
MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Djibouti languages

- 4 languages are spoken in Djibouti
- French and Arabic are official languages
- Somali and Afar are natives and widely spoken

LABORATOIRE D'INFORMATIQUE CERI
339 Chemin des Meinajariès
BP 1228
84911 AVIGNON CEDEX 09
Tél. + 33 (0)4 90 84 35 09
Fax. + 33 (0)4 90 84 35 01
secretariat-lia@univ-avignon.fr
http://www.lia.univ-avignon.fr

Laboratoire d'Informatique
Université d'Avignon

UNIVERSITÉ D'AVIGNON
MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Somali language

- 15 millions of speakers
- Djibouti, Somalia, Ethiopia and Kenya
- Cushitic language within the Afro-asiatic family

LABORATOIRE D'INFORMATIQUE CERI
339 Chemin des Meinajariès
BP 1228
84911 AVIGNON CEDEX 09
Tél. + 33 (0)4 90 84 35 09
Fax. + 33 (0)4 90 84 35 01
secretariat-lia@univ-avignon.fr
http://www.lia.univ-avignon.fr

Laboratoire d'Informatique
Université d'Avignon

UNIVERSITÉ D'AVIGNON
MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Somali language

- 22 consonants and 20 vowels
- Tone accent language
- Written system uses Roman letters (since 1972)
- Words are composed by the concatenation of sub-words/syllables (called "Roots" in our work).

Examples :

Birlab → Bir (CVC) lab (CVC)
Agoon → Ag (VC) oon (VVC)
Ugub → Ug (VC) ub (VC)

LABORATOIRE D'INFORMATIQUE CERI
339 Chemin des Meinajariès
BP 1228
84911 AVIGNON CEDEX 09
Tél. + 33 (0)4 90 84 35 09
Fax. + 33 (0)4 90 84 35 01
secretariat-lia@univ-avignon.fr
http://www.lia.univ-avignon.fr
Laboratoire d'Informatique
Université d'Avignon

UNIVERSITÉ D'AVIGNON

MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Corpus

No corpus available for Somali language

- Collect text from internet (text)
- Record speech (audio)

LABORATOIRE D'INFORMATIQUE CERI
339 Chemin des Meinajariès
BP 1228
84911 AVIGNON CEDEX 09
Tél. + 33 (0)4 90 84 35 09
Fax. + 33 (0)4 90 84 35 01
secretariat-lia@univ-avignon.fr
http://www.lia.univ-avignon.fr

Laboratoire d'Informatique
Université d'Avignon

UNIVERSITÉ D'AVIGNON

MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Text Corpus

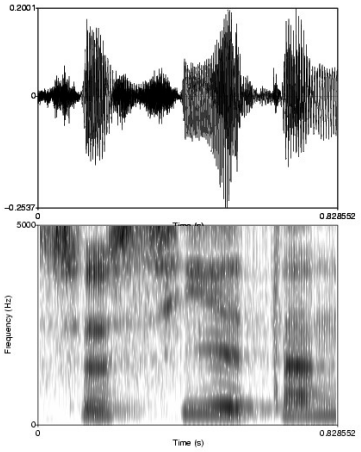
- Only newspapers on internet are available
- WARGEYS corpus : from newspapers (2002-2004)
- 3 millions of words downloaded from internet (broadcast news)
- 121 k different words

LABORATOIRE D'INFORMATIQUE CERI
339 Chemin des Meinajariès
BP 1228
84911 AVIGNON CEDEX 09
Tél. + 33 (0)4 90 84 35 09
Fax. + 33 (0)4 90 84 35 01
secretariat-lia@univ-avignon.fr
http://www.lia.univ-avignon.fr

Laboratoire d'Informatique
Université d'Avignon

UNIVERSITÉ D'AVIGNON
MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Audio Corpus



- **ASAAS**
  - read speech
  - somali news from 2002-2004
    - 10 hours
    - 10 speakers
  - training **ASAAS-train** (8 hours and 57 minutes) ;
  - test **ASAAS-test** (1 hour and 29 minutes) ;

- **RTD**
  - extract of the Djibouti oral tradition
  - 1 hour of speech
  - 5 speakers

LABORATOIRE D'INFORMATIQUE CERI
339 Chemin des Meinajariès
BP 1228
84911 AVIGNON CEDEX 09
Tél. + 33 (0)4 90 84 35 09
Fax. + 33 (0)4 90 84 35 01
secretariat-lia@univ-avignon.fr
http://www.lia.univ-avignon.fr

Laboratoire d'Informatique
Université d'Avignon

UNIVERSITÉ D'AVIGNON
MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

LABORATOIRE D'INFORMATIQUE CERI
339 Chemin des Meinajariès
BP 1228
84911 AVIGNON CEDEX 09
Tél. + 33 (0)4 90 84 35 09
Fax. + 33 (0)4 90 84 35 01
secretariat-lia@univ-avignon.fr
http://www.lia.univ-avignon.fr

Laboratoire d'Informatique
Université d'Avignon

Speech Recognition System

- Acoustic Models
  - Hidden Markov Models
  - Non contextual
  - 20 Vowels
  - 22 Consonants

UNIVERSITÉ D'AVIGNON

MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Speech Recognition System

- Language Model
  - Trigram Language Model
  - 20k words lexicon
  - OOV = 4,90% (Out Of Vocabulary)
  - 723 k bigrams
  - 1 750 k trigrams

LABORATOIRE D'INFORMATIQUE CERI
339 Chemin des Meinajariès
BP 1228
84911 AVIGNON CEDEX 09
Tél. + 33 (0)4 90 84 35 09
Fax. + 33 (0)4 90 84 35 01
secretariat-lia@univ-avignon.fr
http://www.lia.univ-avignon.fr

Laboratoire d'Informatique
Université d'Avignon

UNIVERSITÉ D'AVIGNON
MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

First Experiments

- Speech Recognition System

  LIA continuous speech recognition system : Speeral

Correct	Sub	Del	Ins	Error rate
76.4	20.8	2.8	4.7	**28.2**

LABORATOIRE D'INFORMATIQUE CERI
339 Chemin des Meinajariès
BP 1228
84911 AVIGNON CEDEX 09
Tél. + 33 (0)4 90 84 35 09
Fax. + 33 (0)4 90 84 35 01
secretariat-lia@univ-avignon.fr
http://www.lia.univ-avignon.fr
Laboratoire d'Informatique
Université d'Avignon

UNIVERSITÉ D'AVIGNON
MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

LABORATOIRE D'INFORMATIQUE CERI
339 Chemin des Meinajariès
BP 1228
84911 AVIGNON CEDEX 09
Tél. + 33 (0)4 90 84 35 09
Fax. + 33 (0)4 90 84 35 01
secretariat-lia@univ-avignon.fr
http://www.lia.univ-avignon.fr

Laboratoire d'Informatique
Université d'Avignon

Error Analyses

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**

UNIVERSITÉ D'AVIGNON
MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Standardization problem

- several transcriptions for the same word
- common problem for most of African languages
- Problem of standardization : spelling error
- 3 problems
  - double consonant
    - GU**DD**OOMIYAHA vs. GU**D**OOMIYAHA : director
  - compound-word or not
    - ISGAADSIIN vs. IS GAADSIIN : communication
  - same word written in two different ways
    - W**E**YDIIYAY vs. W**A**YDIIYAY : to ask

LABORATOIRE D'INFORMATIQUE CERI
339 Chemin des Meinajariès
BP 1228
84911 AVIGNON CEDEX 09
Tél. + 33 (0)4 90 84 35 09
Fax. + 33 (0)4 90 84 35 01
secretariat-lia@univ-avignon.fr
http://www.lia.univ-avignon.fr
Laboratoire d'Informatique
Université d'Avignon

UNIVERSITÉ D'AVIGNON
MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Output Standardization

- WER = 21,5%
- relative gain 24%

Standardized corpora	Correct	Sub	Del	Ins	Error rate
none	76.4	20.8	2.8	4.7	**28.2**
ASSAS_test	83.7	14.7	1.6	5.2	**21.5**

LABORATOIRE D'INFORMATIQUE CERI
339 Chemin des Meinajariès
BP 1228
84911 AVIGNON CEDEX 09
Tél. + 33 (0)4 90 84 35 09
Fax. + 33 (0)4 90 84 35 01
secretariat-lia@univ-avignon.fr
http://www.lia.univ-avignon.fr
Laboratoire d'Informatique Université d'Avignon

UNIVERSITÉ D'AVIGNON
MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

LM Standardization

- WER = 20,2%
- relative gain 28%

Standardized corpora	Correct	Sub	Del	Ins	Error rate
none	76.4	20.8	2.8	4.7	**28.2**
ASSAS_test	83.7	14.7	1.6	5.2	**21.5**
WARGEYS	85.1	13.4	1.5	5.3	**20.2**

LABORATOIRE D'INFORMATIQUE CERI
339 Chemin des Meinajariès
BP 1228
84911 AVIGNON CEDEX 09
Tél. + 33 (0)4 90 84 35 09
Fax. + 33 (0)4 90 84 35 01
secretariat-lia@univ-avignon.fr
http://www.lia.univ-avignon.fr
Laboratoire d'Informatique
Université d'Avignon

UNIVERSITÉ D'AVIGNON
MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Djibouti Broadcast News Recognition

- Djibouti Radio Archives
  - 1 hour ;
  - 7 803 words (2 378 different words)
  - spontaneous speech ;
  - thematic and temporal mismatch with the WARGEYS corpora ;
- OOV = 12,48%.

	Correct	Sub	Del	Ins	Error rate
Djibouti BN	46.6	46.4	7.0	8.7	**62.1**

LABORATOIRE D'INFORMATIQUE CERI
339 Chemin des Meinajariès
BP 1228
84911 AVIGNON CEDEX 09
Tél. + 33 (0)4 90 84 35 09
Fax. + 33 (0)4 90 84 35 01
secretariat-lia@univ-avignon.fr
http://www.lia.univ-avignon.fr
Laboratoire d'Informatique Université d'Avignon

UNIVERSITÉ D'AVIGNON
MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Syllables decoding

- Lexicon (4 400 syllables)
- Trigram language model :
  - 189 000 bigrams of syllables,
  - 996 000 trigrams of syllables.
- OOV syllables rate (0,03%)

LABORATOIRE D'INFORMATIQUE CERI
339 Chemin des Meinajariès
BP 1228
84911 AVIGNON CEDEX 09
Tél. + 33 (0)4 90 84 35 09
Fax. + 33 (0)4 90 84 35 01
secretariat-lia@univ-avignon.fr
http://www.lia.univ-avignon.fr
Laboratoire d'Informatique
Université d'Avignon

UNIVERSITÉ D'AVIGNON
MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

LABORATOIRE D'INFORMATIQUE
CERI
339 Chemin des Meinajariès
BP 1228
84911 AVIGNON CEDEX 09
Tél. + 33 (0)4 90 84 35 09
Fax. + 33 (0)4 90 84 35 01
secretariat-lia@univ-avignon.fr
http://www.lia.univ-avignon.fr

Laboratoire d'Informatique
Université d'Avignon

Syllables decoding

Syllables (Roots) decoding =>(RER = Root Error Rate)

	Correct	Sub	Del	Ins	RER
Djibouti BN	57.2	32.3	10.5	4.2	**47.0**

OOV Recognized by Syllabic Recognition

asnaamtaasi	as naam ta si
tafaraaruqa	taf ar aar uq a
faaqidi	aq ad i
shiinaha qudhooda	bish iib a qudh ood a
laba dakhare	lab ad a sar e

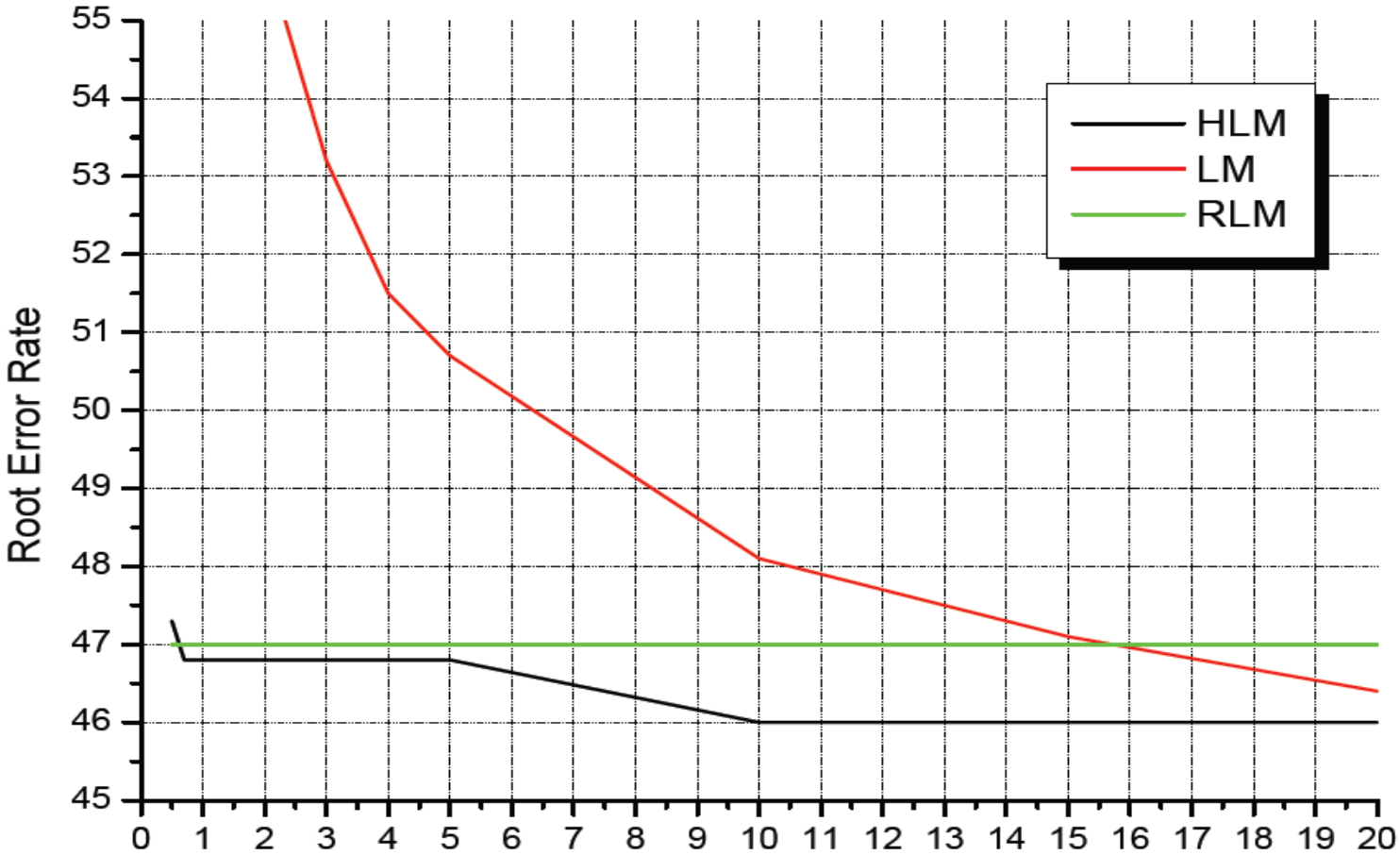
UNIVERSITÉ D'AVIGNON
MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Hybrid Decoding

- Take into account n most frequent words
- Others words are transformed in syllables
- WARGEYS corpora is also transformed in words and syllables
- Hybrid language model (with lexicon n) HLMn
- No difference between syllables and words
- Syllables Results is used

LABORATOIRE D'INFORMATIQUE CERI
339 Chemin des Meinajariès
BP 1228
84911 AVIGNON CEDEX 09
Tél. + 33 (0)4 90 84 35 09
Fax. + 33 (0)4 90 84 35 01
secretariat-lia@univ-avignon.fr
http://www.lia.univ-avignon.fr
Laboratoire d'Informatique
Université d'Avignon

Hybrid Decoding



UNIVERSITÉ D'AVIGNON

MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

LABORATOIRE D'INFORMATIQUE CERI
339 Chemin des Meinajariès
BP 1228
84911 AVIGNON CEDEX 09
Tél. + 33 (0)4 90 84 35 09
Fax. + 33 (0)4 90 84 35 01
secretariat-lia@univ-avignon.fr
http://www.lia.univ-avignon.fr

Laboratoire d'Informatique
Université d'Avignon

UNIVERSITÉ D'AVIGNON
MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Conclusion

- Large vocabulary recognition system for Somali language
- problem of standardization
  - WER relative gain = 28%
- oral patrimony archives ≠ training data
  - thematic and temporal mismatch
- syllables and hybrid decoding seems more appropriate than words

LABORATOIRE D'INFORMATIQUE CERI
339 Chemin des Meinajariès
BP 1228
84911 AVIGNON CEDEX 09
Tél. + 33 (0)4 90 84 35 09
Fax. + 33 (0)4 90 84 35 01
secretariat-lia@univ-avignon.fr
http://www.lia.univ-avignon.fr

Laboratoire d'Informatique
Université d'Avignon