

1 Equipe GETALP - Axe Traitement de Données et de Connaissances à Grande Echelle

1.1 Scientific Presentation

1.1.1 In short

GETALP (Study Group for Machine Translation and Automated Processing of Languages and Speech) was born in 2007 when LIG was created. Born from the virtuous union of researchers in spoken and written language processing, GETALP is a multidisciplinary group (computer scientists, linguists, phoneticians, translators and signal processing specialists...) whose objective is to address all theoretical, methodological and practical aspects of multilingual communication and multilingual (written or spoken) information processing. GETALP's methodology relies on continuous investigations between data collection, fundamental research, development of systems, applications and experimental evaluations.

1.1.2 Research topics

We list below GETALP's areas of research. They find direct applications in various fields such as information access, robotics, assistive technologies and language learning.

- Computer Aided Translation
- Automatic Speech transcription and translation¹
- Processing of under-resourced languages
- Processing / analysis of speech and interactions in ambient environments
- Modelling social affects
- Collection and interoperability of multilingual lexical resources
- Automatic and interactive processes for meaning clarification
- Software engineering for multilingualism
- Automatic summarization of ambient data

1.1.3 Members

GETALP currently includes 17 permanent staff members (2 PR, 11 MC, 1 CR and 3 IR). Over the 2009-2014 period, the team registered 4 arrivals (S. Rossato in 2009, B. Lecouteux in 2011, M. Mangeot in 2011 and V. Aubergé in 2012) and 2 departures (B. Bigi in 2009 and G. Fafiotte in 2012). Table 1 gives the list of current GETALP members.

1.2 Scientific and Technological Results

1.2.1 Common methodology and principles

Ecological approach. One of the specificities of GETALP is the willingness to address the diversity of situations of written or spoken language usage: multiplicity of languages, speakers, dialects, cultures, social contexts and applications, with a special interest for “long tails” (under-resourced languages, atypical speakers ...).

Agnostic approach. The history of the team, and its different scientific cultures, allow to synergize expert and empirical methods, large-scale analysis (big data) and analysis of phenomena requiring fine annotations (beautiful data), induction and models, etc..

¹GETALP members participated in the first international speech translation project CSTAR-2, with the first ever multilingual international demo on 19 July 1999.

Table 1: Member List

Name	Surname	Function	Institution / Section	Date of arrival at LIG
Aubergé	Véronique	CR	CNRS / INSHS34	Oct-12
Bellynck	Valerie	MC	G-INP / 27	September-05
Besacier	Laurent	PR	UJF / 27	September-99 (PR since 2009)
Blanchon	Herve	MC, HDR	UPMF / 27	September-89
Boitet	Christian	PR	UJF / 27	Oct-70
Brunet-Manquat	Francis	MC	UPMF / 27	September-07
Durand	Jean-Claude	ITA IR	CNRS / 07 INS2I	jan-00
Esperança-Rodier	Emmanuelle	MC	UJF / 07	jan-05
Goulian	Jerôme	MC	UPMF / 27	September-03
Guilbaud	Jean-Philippe	IR ITA (50%)	CNRS / 07 INS2I	jan-78
Lecouteux	Benjamin	MC	UPMF / 27	September-11
Mangeot	Mathieu	MC	U-Savoie / 27	September-11
Portet	François	MC	G-INP / 27	Oct-08
Rossato	Solange	MC	Université Stendhal / 07	Jun-09
Schwab	Didier	MC	UPMF / 27	September 08
Sérasset	Gilles	MC	UJF / 27	September-90
Vacher	Michel	ITA IRHC, HDR	CNRS / 07 INS2I	jan-01

Human in the loop. To assist humans in communicative situations, it is necessary to include them in the automated processes (semi-supervised approaches, collaborative and interactive approaches, analysis of system errors).

Production of tools and resources. GETALP develops and distributes free tools and resources (web platform for the cooperative development of multilingual lexical databases, collection of written and oral corpora for processing under-resourced languages, collection of multimodal corpora in contexts of interaction, collaborative system of post-editing and evaluation of machine translations, language-oriented web services, etc.).

A common methodology. common concepts and methods are shared by all team members ; this is especially true for data collection methodology, evaluation of systems (participation to shared tasks), and work with industrial partners on real systems design.

1.2.2 Major Results

Computer Aided Translation (CAT) GETALP has inherited a long tradition of research on various “expert” paradigms for CAT. In the framework of the ANR Traouiero project, all MT systems and modules produced in the past have been “operationalized” through the “Heloise” version of Ariane-G5 produced by V. Berment (thesis in 2004 - associated to GETALP since). A website for the contributive development of MT systems, Lingwarium was launched in 2013. It is based on the (Ariane / Heloise) tools and on the expert methodology of the lab and is meant to be an alternative to Apertium, which is intended for very close languages pairs.

Moreover, since 2007, research on “empirical” approaches (mostly statistical) are also very active. GETALP regularly participates in international MT evaluation campaigns (IWSLT, WMT) and has focused during the last years on confidence and quality measures for machine translation systems. For this, we exploit human post- editions and annotations to build, by learning, estimators based on multiple parameters (words and syntactic tags in source and target languages, alignment information, dependency graphs, etc.). A remarkable recent result by GETALP is a first place in the word-level confidence estimation (WCE) task at the international evaluation campaign WMT 2013. Such systems are efficient and allow us to envisage new generation of interactive machine translation, with a real-time user feedback loop.

On the other hand, the iMAG/SECTra software which provides access in N language to Web pages, with automatic incremental improvement of the pretranslations by contributive human post-editing, was improved and used for many projects: translation of articles of the EOLSS encyclopedia, multilingual access to dozens of web sites. It is a new “paradigm”, replacing diffusion with editorial responsibility by contributive multilingual access.

Finally, we can also mention a strong activity around the construction of resources for machine translation (extraction of polylexical expressions, production of parallel corpora by post-editing or from comparable corpora, building of a very large lexical database linking various languages and UNL); as well as around French \leftrightarrow Chinese MT (2 CIFRE PhD with L&M and a project with YD. Chen from Xiamen University for financial and economic web sites), which uses a hybrid approach (expert segmentation / morphological analysis, empirical alignment and decoding).

Research on this axis is often carried out in collaboration with industry and with foreign labs. Since 2009, 11 PhD theses, including 3 CIFRE were defended on these topics and 5 are ongoing, including 3 CIFRE (see dedicated section on defended and ongoing PhD theses).

Automatic speech transcription and translation GETALP keeps being an international actor in the speech translation domain. A remarkable result is the participation of the group, with very good performance, at the annual IWSLT evaluation campaign (each year from 2009 to 2012 - best result obtained by LIG/GETALP in 2010 on the seminar translation task). On this topic, GETALP has contributed to the problem of efficient coupling between automatic speech recognition (ASR) and machine translation (MT) modules. This effort, as well as advanced use of open-source tools such as Moses (for MT) and Kaldi (for ASR) also allowed the team to develop state-of-the-art systems for different languages and language pairs (Arabic-English, French-English, English-French) providing a credible experimental platform to power research on topics such as: driven decoding for machine translation (text or speech translation), use of MT for cross-lingual speech understanding, study on better handling of multi word expressions (MWE) in MT systems, or other issues mentioned below (under-resourced languages, ASR in smart environments, etc.).

Processing of under-resourced languages This topic has been launched by GETALP more than 10 years ago and remains an area of excellence of the team: 3 journal articles (Speech Communication Journal) in automatic speech recognition for under-resourced languages have recently been published (fast portability of ASR systems with few resources, ASR of languages with special phonological characteristics, etc..). GETALP attracts, on this topic, many high-level foreign students to work on their mother tongue, thus consolidating our network of contacts on all continents (including Africa and South-East Asia). Issues around this axis are related to cultural heritage (defense of linguistic diversity), to global security and to economic development (recent start of the ANR project ALFFA on the development of micro applications for voice interaction on mobile phones in Africa). Many resources (lexical, speech corpora, parallel corpora for under-resourced language pairs) were also collected. Finally, the team is very active in the development and structuring of an international research community around this topic (several workshops were organized, special issues of conferences or journals, etc..).

Processing / analysis of speech and interactions in ambient environments GETALP is active since 2000 on this topic which places speech processing in ambient intelligence (smart homes, smartphones, and more recently companion robots ...). The originality of the team lies in the study of systems coupling low level sensor networks (cheap but semantically poor) with microphones (affordable and semantically rich) sensors rarely used in this field. Since 2009, GETALP coordinated an ANR VERSO project (Sweet-Home) and participated in a TECSAN ANR (Cirido) on the theme of context sensitive voice command for elderly and disabled people. GETALP is also involved in the INTERABOT project (Investissements d’Avenir call) in which communicative primitives of socio-emotional relationship are studied, which establishes a companion robot in a role of socio-relational coach. The team is also working with STMicroelectronics to develop smartphones that can recognize the user context. We were able to make significant contributions on multi-channel automatic speech recognition (ASR) with distant microphones, and on ASR for aged voices as well as activity / scenes recognition and decision making from uncertain data. Another remarkable aspect of this axis is a strong activity in multimodal corpus collection in LIG’s smart home (DOMUS) or outside with respect to ethical rules. So far, we are not aware

of such equivalent corpora. In addition, the team has continued to develop in-house real-time sound analysis tools (PATSH and CirdoX software) which allows us to attract industrial partners. Finally, the team is very active on this topic in both national and international communities and regularly organizes workshops, special sessions and issues.

Modeling social affects GETALP is very much involved in the emerging theme of “Social Affective Speech Signals” (GETALP organized a workshop on this topic - WASSS for Interspeech 2013 – and a special session at Speech Prosody 2014): (1) the theoretical paradigm gives non lexical speech items (prosody, mouth noises, micro facial expressions ...) the role of architect of situated language communication in which the “socio-emotional glue” between speakers supports the semantic exchange (2) the methodological paradigm is based on the theory-experiment-technology loop that leads to collect highly controlled but completely spontaneous corpora (unexpected events) of socio-emotional signals - note that the corpora and the software (Emoz, EEE, HireBot) are developed through several projects (Interabot, NSFC, Innovalangues) (3) this technological paradigm leads to a new generation of dialog systems based primarily on the dynamic exchange of socio-emotional primitives for interaction, the preferred application being the companion robotics.

Structuring and interoperability of lexical resources Concerning structuring and interoperability of multilingual lexical resources, LIG/GETALP seeks to produce large-scale resources. Besides continuing to work on the Jibiki platform [8] which implements an interoperability model centered around XML and a general model of lexical data (Common Dictionary Markup), GETALP has recently become interested in the design of an interoperability model built around the Semantic Web. In this context, a lexical resource called DBnary ⁽²⁾ was built. It is an extraction in RDF of the lexical data of 10 editions of Wiktionary (German, English, Finnish, French, Greek, Italian, Japanese, Portuguese, Russian and Turkish). This resource now has over 39M triplets describing the input and direction of the extracted 12 languages, and more than 3.2 million translations from these 12 languages to more than 1000 target languages. It is accessible in Linked Open Data. DBnary has been awarded the Monnet Challenge prize, an international competition rewarding the best lexical resource based on the LEMON standard. This resource will provide the “vertebral spine” that will aggregate other resources (including those of the DILAF project during which bilingual dictionaries, from 5 African languages into French, were computerized). Such lexical resources are necessary for several NLP tasks; one of them is detailed in the next section.

Automatic and interactive processes for meaning clarification Meaning clarification includes word sense disambiguation (WSD) and is very important for several NLP tasks including machine translation. This axis has emerged at GETALP during last four years under the impulse of two researchers (D. Schwab and J. Goulián) and focuses on multilingual lexical disambiguation and its applications with a particular focus on the enrichment and exploitation of multilingual resources and on multilingual access with guaranteed meaning. This line of research now involves five associate professors (among them 1 HDR, pioneer of the domain in his 1994 PhD) and PhD students and has received funding from two ANR projects (Traouiero and VIDEOSENSE) as well as from Grenoble-2 University. Work on this subject is illustrated by several recent publications and a noted participation to the SemEval evaluation campaign in 2013.

1.2.3 Summary of Publications

Table 2 below summarizes the publications of the team. The complete updated list is available on GETALP publications website ³.

²<http://kaiko.getalp.org/about-dbnary/>

³http://hal.archives-ouvertes.fr/LIG_TDCGE_GETALP/

Table 2: Publications

Year	# All publications	# International journals
2009	59	5
2010	55	6
2011	45	4
2012	88	8
2013	57	10
2014	36	9
Total	340	42

1.3 Visibility and attractivity

1.3.1 Local, national and international ecosystem

GETALP has a rich local ecosystem and is working with other groups of LIG (MRIM, AMA, IIHM, EXMO, PRIMA, MAGMA) and with other Grenoble laboratories. GETALP is an internationally recognized player in the natural language and speech processing communities (many collaborative and industrial projects, organization of JEP-TALN-RECITAL in 2012, etc.). Finally, a remarkable feature is its international network of collaborations and contacts on all continents, which makes GETALP a particularly relevant and convincing stakeholder to contribute to the theoretical and technological Grail of multilingualism. These include the following international collaborations:

- **India.** *IITB (Bombay)*: project Arcus-1, codirection of Indian Master students and work with the team of Prof. P. Bhattacharyya around UNL and processing of Hindi. 3 stays of 2 months as a visiting professor Prof. Bhattacharyya, 1 reciprocal stay by Ch. Boitet. Common organization of COLING-2012 in India. Participation in the ARP UJF-IRD-India. *Pondicherry University*: collaboration with the French department (Prof. Pannirselvame) and invitation K. Vijayanand 1 month (MOU).
- **China.** GETALP is responsible for an international team (9 laboratories, 6 nations) in a project funded by the National Social Science Fund of China on second language learning. In addition, the team is active on French ↔ Chinese MT and currently hosts Y.D. Chen of Xiamen University.
- **South-East Asia. Malaysia.** Close-links with Malaysia, especially with USM (Universiti Sains Malaysia, Penang) for many years, and right now hosting a young professor (T.P. Tan, PhD in Grenoble in 2008). Sabbatical and ongoing thesis of SF Juan (UNIMAS, Sarawak), co-organizing the workshop LMS 2010 with USM. **Singapore.** Franco-Singaporean project (Merlion) on the automatic recognition of multilingual speech with Prof. Haizhou Li (Institute for Infocomm Research, Singapore) and respective visits or exchanges of students and / or postdocs over the period 2009-2012. **Vietnam.** Finally, strong ties (collaborative projects, co-supervision of PhD students) are maintained with the UMI MICA (Hanoi, Vietnam).
- **Russia.** IPPI (Moscow): CNRS-RAS project No. 24179 of 4 years (2010-13) with 30 days scientific stays on each side (cofunded by an ANR project) on MT lingware and software engineering (ETAP-3, Ariane-G5), a project on integration of IPPI tools to build heterogeneous systems, and cooperation on the construction of a large lexical database with UW++.
- **Brazil.** As part of a scientific partnership between the LIG and UFRGS (Porto Alegre), an international laboratory associated with CNRS was created (LICIA). In this context, GETALP carries the CAMELEON project whose objective is to develop collaborative methods for multilingualizing lexicons and ontologies (2011-2014, funded by CAPES-COFECUB). Very recently, the AIM-WEST project (analysis and integration of polylexical expressions for improving machine translation of text and speech) involving CNRS/INS2i, FAPERGS (Rio Grande do Sul) and FAPESP (São Paulo), driven by GETALP, was accepted. Finally, close collaboration between GETALP and the University of Ouro Preto allows the exchange of researchers on the topic of emotional speech.

- **Japan.** Collaboration with Japan, historically dense, has been maintained in the past 5 years, particularly with the National Institute of Informatics, University of Tokyo, and Waseda University, in particular through 3 PhD co-directions (M. Daoud, F. Cromières, Y. Sasa) and a bilateral project (Sakura Survitra/JP). M. Mangeot (GETALP) also received a one-year scholarship by the Hosei University in Tokyo from 2014 to 2015.
- **Africa.** Network: highly developed contacts with several countries in sub-Saharan Africa. *Ethiopia:* thesis supervision with University of Addis Ababa, home post-docs for research on Amharic. *Niger and Mali:* bilingual dictionaries between French and several African languages⁴. *Other:* work on speech processing (ASR) for African languages project⁵, involving numerous contacts on the continent.
- **Romania.** Many scientific exchanges with Professor C. Burileanu (Polytechnic Institute of Bucharest, "Human-Computer Dialog Group"), co-supervision of Master students, host lab for a doctoral research fellowship (H. Cucu), joint publications (1 review article, 5 international conferences). GETALP recently hosted a visiting researcher (C. Munteanu) of the institute and a common journal article has been accepted and is under review.

1.3.2 Highlights

- L. Besacier new member of IUF (2012-2017),
- Organization of JEP-TALN-RECITAL (major francophone international conference on Speech and NLP with 350 people gathered in Grenoble in June 2012),
- Ch. Boitet co-chair of COLING-2012 (major international conference on computational linguistics),
- PhD thesis award (delivered by AFCP - Francophone Speech Communication Association) obtained by J. Kahn in 2012,
- G. Sérasset won, with DBnary, the Monnet Challenge in 2012 (on lexical linked data resources), as well as the Lexical Linked Data Challenge in 2014 (during LREC 2014)
- Ch.Boitet co-chair of the COST MUMIA action (2011-14),
- Best Paper Award (Young Investigator) for M. Verdurand supervised by S. Rossato (International Conference on Stuttering, Rome, Italy, 2012),
- Best Paper Award for V. Auberge and co-authors (Workshop on Emotion, Social Signals, Sentiment and Linked Open Data, during LREC 2014),

1.3.3 Committees, Expertise, Recruitment

GETALP members are regularly in ANR evaluation committees (G. Sérasset), expertise projects submitted to FUI and ANR calls (Besacier, Blanchon, Boitet Portet, Aubergé, Rossato), EU ERC (Besacier), Quebec research council (Portet, Vacher), ARC6 of Rhône-Alpes region (Besacier). Moreover, they are also asked to participate in PhD or HDR committees (Avignon, Orsay, Nancy, Le Mans, Lille, Montpellier, Toulouse, UPMC, EC Lyon, Telecom-Paritech, Aix-Marseille, IIT Bombay, Bucharest Polytechnic, USM in Malaysia). L. Besacier is also a member of the scientific committee of the ADM (Advanced Data Mining) action of the Grenoble Persyval labex and participated in the scientific committee of the inaugural Jean Kuntzmann prize.

1.3.4 Editorial Work (journal, conferences)

Conferences and Workshops

- Co-chair of the SLSP 2014 conference (2nd International Conference on Statistical Language and Speech Processing)

⁴<http://www.dilaf.org>

⁵<http://alffa.imag.fr>

- Chair of the Workshop on South and Southeast Asian Natural Language Processing (WSSANLP 2014) associated with COLING 2014,
- Participation in the organizing committee of Interspeech 2013 Lyon (1500 persons - coordination of satellite events)
- Chair of the jury for the Google-Show-and-Tell-Award (for Interspeech 2013)
- Co-organizing a special session at Interspeech 2011 (Speech technology for under-resourced languages) and Interspeech 2014 (Speech Technology for Ambient Assisted Living)
- Organization of a special session at EUSIPCO 2012 (Audio analysis in Smart Homes)
- Chair and organization of the first two and the fourth international SLTU workshops (Spoken Language Technologies for Under-resourced Languages) and permanent member of the SLTU board,
- Chair and organization of 4th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT 2013), Satellite of InterSpeech2013
- Chair and organization of the 1st Workshop on Social Affective Speech Signals (WASSS 2013), Satellite of InterSpeech2013
- Participation to reviewing or program committees for international conferences: Interspeech, IEEE ICASSP, IEEE ASRU, IEEE SPeD, ACL, EAACL, COGALEX, COLING, CORIA, EUSIPCO, Speaker Odyssey, IWSLT, EAMT, LREC, NAACL-HLT, IEEE/ACL SLT, Speech Prosody, IJCAI, TALN, JEP, RECITAL, AIME, MEDINFO, MIE, ISG, SLPAT etc.

Journals

- Editor-in-Chief for a special issue of the Speech Communication journal (Elsevier) 2014 (special session on "Speech technology for under-resourced languages").
- Member of the editorial board of the French TALN journal (Natural Language Processing) since 2011
- Guest Editor for a Special Issue on Spoken Language Processing for the TAL journal (2014)
- Guest Editor for Special Issue on speech processing technologies for the ACM TACCESS journal (2014)
- Member of the editorial board of the Journal of Ambient Intelligence and Smart Environments (JAISE) since 2013
- Proofreading of articles for international journals IEEE/ACM Transactions on Acoustics, Speech and Language Processing (IEEE/ACM ASL) ; Computer Speech and Language Journal ; Speech Communication Journal ; IEEE Transactions on Speech and Audio Processing ; IEEE Signal Processing Letters ; IEEE Transactions on Signal Processing ; IEEE Transactions on Multimedia ; IEEE Transactions on Information Forensics and Security ; IEEE Transactions on Systems, Man, and Cybernetics; EURASIP Journal on Audio, Speech, and Music Processing; Pattern Recognition Letters ; Machine Translation Journal ; Language Resources and Evaluation Journal (LRE) ; JoSS ; Frontiers in Emotion Science ; Methods of Information in Medicine; Artificial Intelligence in Medicine; International Journal of Adaptive Control and Signal Processing ; Computer Methods and Programs in Biomedicine; Pervasive and Mobile Computing; T-ASE; Sensors;

1.4 Social, economical, and cultural impact

1.4.1 Current Contracts with Industrial Partners

- ALFFA : African Languages in the Field - Fundamentals and Automation - ANR 136K€ - with Voxygen SA.

- CAMOMILE: Collaborative annotation of multi-modal, multi-lingual and multi-media documents - ANR 250K€ - with Vocapia Research.
- CIRDO: Compagnon Intelligent Réagissant au Doigt et à l’Oeil (Intelligent Companion Obeying at User’s Beck and Call) - ANR 210K€ - with Technosens SAS
- AXiMAG: Interactive Multilingual Access Gateway (young spin-off).
- INTERABOT: Interactions naturelles avec un robot compagnon (Natural interactions with a companion robot), IA (Investissements d’Avenir) 267K€ with Awabot.
- PEA-TRAD: Traduction pour l’aide à l’analyse documentaire (Translation for Documentary Analysis) - PEA DGA 50K€ - with Airbus Defence and Space (ex Cassidian).
- QCOMPERE: Consortium Quaero pour la Reconnaissance Multimodale des Personnes (Quaero Consortium of Multimodal Person Recognition) - ANR - 81K€ - with Vocapia Research.
- CIFRE PhDs: 4 PhD students are currently funded by a CIFRE grant: 2 with Lingua & Machina, 1 with ST-Microelectronics, 1 with Object Direct (Viséo).
- Completed projects over the period 2009-2014: ANR OMNIA (with Xerox RCE), ANR VIDEOSENSE (with Ghanni SA)


1.4.2 Patents

- Filing patent application with ST-Microelectronics “Loudness adaptation at audio rendering of an audio signal” (D. Blachon, PhD CIFRE LIG-ST & S. Tassard ST).
- 3 APP deposits done during the ANR Traouiéro project.

1.4.3 Interdisciplinary Aspects

The dynamics of GETALP is notably due to the complementarity of its members, enabling an “in depth” approach from data collection to evaluation, from understanding of fundamental communicative phenomena to industrial applications.

1.4.4 Popularization for the General Public

- **France 5** documentary “Défense de sourire” (“Smile Forbidden”), broadcast on several French TV channels, 2013.
- **M6** “E = M6” replay 2012 2013 (emotion detector).
- **France 5** program on the polygraph (lie detector) “On n’est pas que des cobayes” 2012, replay in 2013.
- **France Inter** “Les passagers de la nuit” (“The night passengers”, about sounds in the city), in 2011, 2012 and 2013.
- **RTL** “Les grosses têtes”, on the polygraph (lie detector), 2010, replay in 2012.
- **France Culture** “Science publique” (“Science made public”) on speech synthesis, 2007, replay 2008, 2009, 2010, 2013.
- Journal article on machine translation in **Les dossiers de la recherche** (No. 4, June-July 2013, pp. 88-89 and pp. 96-97).
- **Forum-4i**  - Technology for health: the economic and social issues, convention center WTC Grenoble, May 2009.

1.5 Team organization and life

1.5.1 Seminars and scientific life

Scientific life The GETALP team is housed in the IM2AG B building and spreads over two floors (second and third). Permanent researchers meet about every three weeks to formally discuss all topics related to the group scientific life: research, administration, project submissions, student (Master, PhD) subjects, etc. The team maintains a website⁶ as well as an intranet. The aim of this web site is to be a showcase on the activities of the research group. It is made using a collaborative editing tool (wiki style). Every Tuesday morning, a more informal meeting is held in the “GETALP café”, with the aim to share ideas in a friendly atmosphere. Technical meetings are also held each week (chaired by L. Besacier) so that the various team members can share their current work / problems.

GETALP seminars Team seminars are regularly scheduled on Thursday afternoon, alternating with team meetings. The following list gives examples of speakers invited recently, since 2009: Tom Richens, Violeta Seretan, Patrick Drouin, Aline Villavicencio, Albert Gatt, Igor Boguslavski, Shinsuke Mori, Ehud Reiter, Pushpak Bhattacharyya, Renata Viera, Tanja Schultz, Yuya Akita, Toshiaki Nakazawa, Yves Lepage, Martin Kay, Karen Fort, Lucia Specia, Hansjoerg Mixdorff, Yidong Chen, Emmanuel Vincent. A two-day green seminar was also organized with LIDILEM laboratory in July 2013. An important outcome of it lies in the social exchanges it created and in the cohesion it provided to GETALP team (as well as new collaborations with LIDILEM).

1.5.2 Taking Into Account the Recommendations of the Previous Evaluation

We summarize below some comments taken from the previous AERES assessment of the team and respond to them briefly.

AERES “Some themes could be combined, rearranged. Theoretical questions that the team intends to solve should be highlighted, in particular, those related to hybridization methods for MT. We recommend both a reduction in the number of research themes and their redefinition to put more emphasis on scientific issues and less on applications. ”

GETALP As illustrated by this document (as well as by the website of the team), we have deeply reorganized the research topics of GETALP. The methodology and positioning of the group are more specifically described in the previous sections of this document. All 9 research axes (for 17 permanent members) contribute to the Grail of the team, that consists in addressing all theoretical, methodological and practical aspects of multilingual communication and information processing.

AERES ”The team suffered the departure of one DR and one CR and hosts now only a CR at 40 %. This CR should therefore be associated at 100 % to partially fill this gap.

GETALP This has been done

AERES The arrival of new academic staff should above all consolidate these research topics and not add any dispersion.

GETALP Recruitment of B.Lecouteux to consolidate the aspects of empirical translation / automatic speech recognition; arrival of Mr. Mangeot on the subject of lexical resources and processing of under-resourced languages; arrival of V. Aubergé and S. Rossato to reinforce the oral / speech part.

AERES ”Over the period 2005-2008, the level of publications remains medium in terms of the size of the team: some national and international journals (12) and few conferences of A rank.”

GETALP The number of publications in international journals has increased over the last four years (28 journals in 2010-2013 against 18 in 2007-2010).

AERES ”Too long PhD period for one quarter of the students’

⁶<http://getalp.imag.fr/xwiki/bin/view/Main/>

GETALP This was related to some particular cases in 2009-2010 and since then the average length of PhDs is 41 months for the period 2010-2013 (equivalent to the average PhD duration at ED MSTII).

AERES "Spreading (essaimage) is only effective abroad with the recruitment of foreign doctors as lecturers or researchers in their country of origin"

GETALP Two doctors were recruited in 2013 as permanent associate professors in French universities (MCF - Marseille and Avignon), 1 doctor is at the JRC-EU; 1 at LNE, etc.

1.6 Training through research, educational involvement

1.6.1 GETALP's PhDs: Current and Awarded

The GETALP team currently has **17 PhD theses in progress**. The detailed topics and supervisors are available⁷. The detailed list shows that 13 researchers or (associate or full) professors of GETALP are involved in at least one co-supervision (87% of permanent members co-supervise a thesis).

Most students (16/17) are enrolled in a graduate school of the University of Grenoble, mainly the doctoral school called *Mathematics, Science and Information Technology, Computer Science* (MSTII - 14 students), but also the doctoral school for Languages, Literature and Humanities (LLSH - 2 students).

The list below shows the **19 PhD theses and 1 HDR recently defended during the period 2009-2014** (as well as the current situation of graduated students). The average duration of PhDs carried out at GETALP is equal to the average duration of PhDs at the MSTII doctoral school (around 41 months).

2013

- 1 P. Chahuara. Contrôle intelligent de la domotique à partir d'informations temporelles multiresource imprécises et incertaines. Nouvelle thèse, Université de Grenoble, Grenoble, France, March 2013. P. Chahuara is postdoc at EU Joint Research Center (JRC) in Italy.
- 2 J. Poignant. Identification non-supervisée de personnes dans les flux télévisés. Nouvelle thèse, Thèse en informatique à l'université de Grenoble, 2013. J. Poignant is ATER at UPMF.
- 3 M. Potet. Vers l'intégration de post-éditions d'utilisateurs pour améliorer les systèmes de traduction automatique probabilistes. Nouvelle thèse, Université de Grenoble, Grenoble, France, April 2013. M. Potet is currently turning to teaching in secondary education.

2012

- 4 B. Jabaian. Vers une approche conjointe pour la portabilité multilingue d'un système de compréhension de la parole. Nouvelle thèse, Université d'Avignon, Co-supervision LIG-LIA, Dec. 2012. B. Jabaian got a permanent position (MCF) at University of Avignon in 2013.
- 5 C. Ramisch. A generic and open framework for multiword expressions treatment: from acquisition to applications. Nouvelle thèse, University of Grenoble (France) and Federal University of Rio Grande do Sul (Brazil), Grenoble, France, Sept. 2012. C. Ramisch got a permanent position (MCF) at University of Aix-Marseille in 2013.
- 6 D. Rouquet. Multilinguisation d'ontologies dans le cadre de la recherche d'information translingue dans des collections d'images accompagnées de textes spontanés. Nouvelle thèse, Université de Grenoble, Grenoble, France, April 2012. D. Rouquet is postdoc at the LIDILEM laboratory.

⁷<http://getalp.imag.fr/xwiki/bin/view/Projects/ThesesEnCours>

2011

- 7 T. N. D. Do. Extraction de corpus parallèles pour la traduction automatique depuis et vers une langue peu dotée. Nouvelle thèse, Université de Grenoble, Grenoble, France, Apr. 2011. D. Do is permanent associate professor at Hanoi Institute of Technology (Viet-Nam).
- 8 J. Kahn. Parole de locuteur: performance et confiance en identification biométrique vocale. Nouvelle thèse, Université d'Avignon, Co-supervision LIG-LIA, 2011. Prix de thèse AFCP 2011 (Award). J. Kahn is postdoc at LNE.
- 9 S. Sam. Vers une adaptation autonome des modèles acoustiques multilingues pour le traitement automatique de la parole. Nouvelle thèse, Université de Grenoble, Grenoble, France, April 2011. S. Sam is permanent associate professor at Institute of Technology of Cambodia.
- 10 M. Vacher. Analyse sonore et multimodale dans le domaine de l'assistance à domicile. HDR, University of Grenoble, Grenoble, France, Oct. 2011.

2010

- 11 M. Daoud. Usage of non-conventional resources and contributive methods to bridge the terminological gap between languages by developing multilingual preterminologies. Nouvelle thèse, Université Joseph Fourier (UJF), Grenoble, France, Dec. 2010. M. Daoud is permanent associate professor at university of Amman (Jordan).
- 12 A. Fraisse. Localisation interne et en contexte des logiciels commerciaux et libres. Nouvelle thèse, Université Joseph Fourier (UJF), Grenoble, France, 10 juin 2010. A. Fraisse is ATER at University of Paris 11.
- 13 C.-P. Huynh. Des suites de test pour la TA à un système d'exploitation de corpus alignés de documents et métadocuments multilingues, multiannotés et multimodaux. PhD thesis, Université Joseph Fourier (UJF), Grenoble, France, 17 juin 2010. C-P Huynh is the dean of the Faculty of Computer Science at the University of Danang (Vietnam).
- 14 M. G. A. Malik. Méthodes et outils pour des problèmes faibles de traduction (Methods and Tools for Weak Problems of Translation). Nouvelle thèse, Université Joseph Fourier (UJF), Grenoble, France, 9 juillet 2010. A. Malik is permanent associate professor at University King Abdulaziz (Saudi Arabia).
- 15 V. Nikoulina. Modèle de traduction probabiliste à fragments enrichi par la syntaxe. Nouvelle thèse, Université Joseph Fourier (UJF), Lab. LIG, Eq. GETALP, Grenoble, France, March 2010. V. Nikoulina is permanent researcher at Xerox RCE.
- 16 S. Seng. Vers une modélisation statistique multiniveau du langage, application aux langues peu dotées. Nouvelle thèse, Université Joseph Fourier (UJF), March 2010. S. Seng is permanent associate professor at Institute of Technology of Cambodia.
- 17 Fabien Cromières. Vers un plus grand lien entre alignement, segmentation et structure des phrases. Nouvelle thèse, Université Joseph Fourier (UJF), Lab. LIG, Eq. GETALP, Grenoble, France, January 2010. F. Cromières is assistant professor at University of Kyoto (Japan).

2009

- 18 V. Archer. Graphes linguistiques multiniveau pour l'extraction de connaissances : l'exemple des collocations. Nouvelle thèse, Université Joseph Fourier, Grenoble-I, Sept. 2009. V. Archer is research engineer at Gamed S.A (Marseille).
- 19 A. Falaise. Conception et prototypage d'un outil web de médiation et d'aide au dialogue tchaté écrit en langue seconde. Nouvelle thèse, Université Joseph Fourier – Grenoble-I, Sept. 2009. A. Falaise is postdoc at LIDILEM, Université Stendhal — Grenoble-3).
- 20 H.-T. Nguyen. Des systèmes de TA homogènes aux systèmes de TAO hétérogènes. Nouvelle thèse, Université Joseph Fourier, Grenoble-I, Dec. 2009. H-T Nguyen is research engineer at Polyspot S.A. (Paris).

1.6.2 Involvement in research Masters and doctoral schools

Since September 2008, L. Besacier is in charge of the computer science speciality at the graduate school of Mathematics, Information Technology and Informatics (MSTII). The graduate school has 400 students in four specialties, the informatics speciality has about 230 students from 30 different nationalities attached to four universities and seven research laboratories.

Moreover, V. Aubergé is in charge of the TALEP axis of a Master on Language Technologies (in French: Industries des Langues - IdL - rated A+) in which four faculty members of the team are

involved. Finally, GETALP is in charge of the "Speech and Language Engineering" module that opens every 2 years in the international Master of Informatics at Grenoble (MOSIG - rated A+).

1.7 Strategy and Research Project

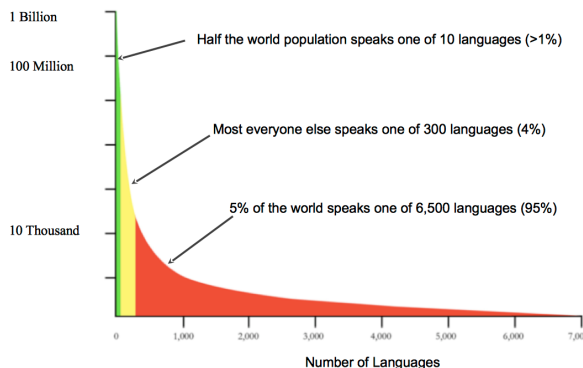


Figure 1: The "long tail" of languages

For the next year, GETALP will continue to contribute to the theoretical and technological **Grail of multilingualism**. Nowadays, there are around 60 languages covered by majors companies of the field (language and speech technologies), but the "long tail" of languages (presented in Figure 1) shows that there is commercial interest in enabling the 300 most widely spoken languages in the digital domain (that represents 95% of humanity). Also, while the other 6,000 languages are not of commercial interest, there are other reasons to enable them if possible: to provide access to information, to provide a critical new domain of use for endangered languages, to foster a better linguistic knowledge of them, to respond in a crisis ("surge languages"), etc. This is a challenge for the coming 20 years and GETALP believes that it can be reached following the group's methodology relying on ecological, agnostic and multidisciplinary approaches with the human placed at the center of the multilingual communication process. We detail below two main challenges that GETALP will focus on for the next years. Some of their aspects involve a deep collaboration between two or more existing research axes (defined in the group presentation earlier) and some other aspects open new avenues of research.

1.7.1 New generation Machine Translation and broad language coverage

In his talk entitled *Five open problems in machine translation* given at EMNLP 2013, A. Lopez from JHU mentioned the following challenges for the coming years: (P1) Translation of low-resource language pairs, (P2) Translation across domains, (P3) Translation of informal text, (P4) Translation into (or from) morphologically rich languages, (P5) Translation of speech. This comforts us strongly in the group's research axes and we will continue to focus on language portability issues (for instance cross-lingual annotation projection) and under-resourced language pairs (P1), extending our international network to collect language resources using innovative approaches (including crowdsourcing and collaborative web). Moreover, as far as high quality translation for dedicated domains (P2) is concerned (huge industrial demand), we believe that it is very important to synergize multilingual lexical resources collection, word-sense disambiguation, efficient user-interfaces (for post-edition, interactive disambiguation, error correction) and machine translation technologies altogether. This will lead to a new generation MT (or CAT) involving an integrated and virtuous loop between humans and machines. As far as hybrid MT is concerned, we adopt a pragmatic approach that involves the use of efficient NLP preprocessors (morphological analyzers – important for (P4), parsers, etc.) for integrating sparse features of different nature (lexical, syntactic, semantic, system-based) into statistics-based translators. Finally, the new challenges for speech-to-speech (S2S) translation (P5) that will be handled at GETALP are: interactive meaning clarification process in S2S, personalization of S2S systems (using cross-lingual speaker conversion), confidence estimation for speech translation and automatic simultaneous translation.

1.7.2 Spontaneous spoken interaction in ambient environments

The local ecosystem of GETALP (other teams at LIG, DOMUS smart home), as well as the group's evolution (arrival of 3 new speech researchers since 2009) comfort us in developing research activities on spontaneous speech interactions for ambient intelligence (smart homes, smartphones, companion robots). We want to propose new generation dialog systems for natural interaction (technological and scientific challenges being distant and multi-channel speech recognition, transcription of spontaneous speech, analysis of paralinguistic traits and modeling of social affects). Privileged applications are companion robots and assistive technologies (as well as new exciting domains such as ambient data summarization). GETALP has recently submitted several projects (ANR, Inv. d'Avenir) on these topics. The challenge of personalization is also very exciting and the complementarity of GETALP's members, as well as previous research results on elderly speech processing makes GETALP a particularly relevant and convincing stakeholder to contribute to speech processing for (specific) speakers such as the elderly, children, etc. All these aspects contribute to redefine human language understanding where the "socio-emotional glue" between speakers (or between speakers and machine) supports the semantic exchange.

1.8 Self assesment

Strengths.

- Three senior researchers internationally recognized in their domain (machine translation, speech recognition and social affective speech)
- Outstanding network of contacts and collaborations on all continents (for research on multilingual and multicultural communication)
- Diversity of scientific cultures and synergy between empirical and experts methods for NLP

Weaknesses.

- Lack of critical mass on topics that become more and more demanding (machine translation)
- Low participation to UE projects

Opportunities.

- Emerging workgroup on social robotics at LIG (and take-off of social robotics at national and international levels)
- Digital humanities as an emerging topic in Grenoble (with University Grenoble 3)
- Under-resourced languages now a hot topic both for e-inclusion and global security (new funding opportunities)

Threats.

- One outstanding professor will retire soon: a new senior professor recruitment will be crucial for the team which now includes 17 permanent staff members and few senior researchers
- Decrease of ANR funds may jeopardize the team overheads: new funding sources search needed (FUI, direct industrial contracts, DGA)