

## US-Morocco: Building a Human Language Technology Research and Education Agenda

Monaco, Soudi, Poley, Lee

### Project Summary

In spring of 2003 Gregory Monaco and Abdelhadi Soudi met at the *Workshop on Information Technology* in Rabat, Morocco, and initiated a dialogue on an interdisciplinary research and education collaboration (linguistics, psychology and computer science). Monaco has returned to the position of Director for Research with the Great Plains Network, a consortium of 14 universities in a seven state region, and Soudi has taken a sabbatical from the *Ecole Nationale de L'Industrie Minérale* (ENIM) in Morocco to serve as an Alexander von Humboldt Research Fellow at the German Research Center for Artificial Intelligence, Language Technology Lab, Saarbrucken, Germany. Both have remained active in redefining the collaboration and engaging others from their respective communities with the goal of creating a *US-Morocco Collaborative Human Language Technology Research and Education Agenda*.

#### Intellectual Merit

The goal of this agenda is to

- Identify a set of common research issues and a joint research plan in Human Language Technology, relevant both to languages and dialects in commercial and educational use (English, Standard Arabic, Moroccan Arabic) and to indigenous languages (e.g., Berber, Native American languages), which are amenable to interdisciplinary analysis (e.g., maximizing human comprehension of machine translated instructions from Standard Arabic to Berber).
- Initiate development of a joint program of study in Human Language Technology, which is interdisciplinary, takes advantage of developments in distance education to initially team scholars from the two countries to develop curricula and to teach together, and involves human resource sharing (students, faculty) between the United States and Morocco.

Initial phases of the project are underway. Funding is sought to bring Abdelhadi Soudi to the United States and the Great Plains Network region for one week (February, 2004) to finalize a research and education agenda for 2004 – 2005 with US collaborators.

#### Broader Impacts

Impacts of the work to be done include

- Cultural: Assisting to safeguard linguistic and cultural diversity in the information society of tomorrow by strengthening the position of Arabic and other local dialects in linguistics and language technology;
- Social-Political: Promoting cooperation between the United States and (North) Africa for the purpose of developing basic components for the multilingual information society;
- Economic: Easing the entrance requirements for United States companies into the Moroccan market and vice versa by providing the basis for automatic translation tools which can be used to translate documentation and marketing material; and,
- Educational: Providing a platform for educational exchange and distance education leading to workforce development and increased economic

US-Morocco: Building a Human Language Technology Research and Education Agenda

Monaco, Soudi, Poley, Lee

opportunity for citizens as well as interdisciplinary, cross-cultural training of Human Language Technology graduate students.

### Project Description

By its very nature, Human Language Technology or HLT<sup>1</sup> is a multidisciplinary<sup>2</sup> enterprise involving, at a minimum, the fields of linguistics, psychology, computer science and (special) education. When taking into consideration the actual domain of information under consideration (e.g., medicine, law, chemistry, physics) additional specialization may be required.

Consider the case of building an effective system for machine translation of text between languages that maximizes reader comprehension. We know that comprehension of information under ideal conditions (two native speakers of the same language, a relaxed setting) rarely, if ever, results in one to one mapping between the speaker's intent and the listener's understanding (Harris & Monaco, 1978). Two variables that affect comprehension of written information are text difficulty and skill of the reader (Wolfe, Schreiner, Rehder, Laham, Foltz, Kintsch, & Landauer, 1998). It is not surprising that, for information that is translated by a computer (Machine Translation or MT) from speech to text or between languages or between sensory modalities, rates of comprehension drops precipitously due to characteristics of the author, accuracy of the computer program which does the translation, and, finally, characteristics of the reader (Murphy, 2000).

The factors involved in end-user comprehension of machine translated text are amenable to interdisciplinary analysis of the translation scheme, itself, for example knowledge-based (Nyberg & Mitamura, 1992; Nirenburg, Carbonell, Tomita, & Goodman, 1996; Soudi, Cavalli-Sforza, & Jamari, 2002) or example-based (Brown, R. 1996), plus an analysis of the host of human information-processing and text-representation variables ultimately involved in human comprehension.

#### Purpose of This Proposal

The purpose of this proposal is to obtain funds to assist in completing the first phase of establishing a common, interdisciplinary research and

---

<sup>1</sup> E.g., machine translation of speech to text and text to text, information retrieval and extraction for specialized knowledge domains, mono- and multilingual information access, document categorization and text summarization tools, language learning and interpretation aides, language augmentation systems for people with special needs

<sup>2</sup> We use the term *multidisciplinary* to describe the fact that the field of Human Language Technology covers many disciplines; we use the term *interdisciplinary* to describe the activity of individuals from diverse disciplinary backgrounds working together to identify and resolve common problems.

education agenda, in the area of Human Language Technology, among multiple institutions in the United States and Morocco. In particular, this collaboration is focusing on the complex problem of meeting the language technology needs of societies with multiple dominant and indigenous languages, as in the United States and Morocco.

The goal of this phase of the project is to

- Identify a set of common interdisciplinary research issues and a joint research plan in Human Language Technology, relevant both to languages in commercial and educational use (English, Standard Arabic, Moroccan Arabic) and to indigenous languages (e.g., maximizing human comprehension of machine translated instructions from Standard Arabic to Berber and vice versa).
- Initiate development of a joint program of study in HLT, which is interdisciplinary, takes advantage of developments in distance education to initially team scholars from the two countries to develop curricula and to teach together, and involves human resource sharing (students, faculty) between the United States and Morocco.

#### Project Timeline

The activities of this project with respect to the research and education agenda are

- Identification of the primary languages of interest;
- Identification and prioritization of potential
  - Research questions
  - Application areas (e.g., machine translation, information extraction, computer-aided instruction)
  - User communities
  - Knowledge-domain area (e.g., legal, medical, behavior analysis)
  - Information representation variables
  - Joint educational projects
- Face-to-face meeting of collaborators to define scope of research, responsibilities and timeline for completion. The collaborators will also develop a plan to undertake one joint educational project.

The timeline for these activities is presented in Table 1 (Activities 1, 2, & 3).

US-Morocco: Building a Human Language Technology Research and Education Agenda

Monaco, Soudi, Poley, Lee

	Activity	Year/Quarter							
		2003		2004				2005	
		3	4	1	2	3	4	1	2
1	Identify primary languages of interest.								
2	Identification and prioritization of potential research questions, application areas, user communities, knowledge-domain area, information representation variables, joint educational projects.								
3	Face-to-face meeting of collaborators to define scope of research, responsibilities and timeline for completion. The collaborators will also develop a plan to undertake one joint educational project.								
4	Research Activities: Develop one full-fledged exemplar application in knowledge domain area of choice (e.g., education, social, economic, medical/pharmaceutical information, legal);								
5	Educational Activity: Develop and Implement Joint Educational Project								
6	Publication and Dissemination of Results								
7	Planning and Evaluation Meeting: Expansion to new knowledge domain areas and new languages (e.g., Native American languages), broaden participation, explore transfer of technology to industry								
Funding is sought for Professor Soudi's travel in the 1st Quarter of 2004 for Activity 3: Face-to-Face Research and Education Planning Meeting. No funds are sought for any other activity. Although the scope of this project ends with Activity 3, the additional plan of activities is included to demonstrate how the proposed project fits into the broader scope of the collaboration.									

While much of the initial work has been and will continue to be accomplished, remotely, via virtual teams, we consider the face-to-face meeting (Table 1, Activity 3) to be vital to the successful launching of the project.

Background of the Collaborators

Monaco's background is in psycholinguistics, transfer of training, and computer application development. He has experience building collaborative teams and extensive experience in project management for applied product development (Monaco & Smith, 1989, 1991; Monaco & Tomiser, 1992, 1995, 1996, 2001, 2002, 2003; Monaco & Wu, 1999). Monaco will coordinate the efforts of the US team.

Dr. Lee has extensive experience in information retrieval (Lee, in press; Lee, submitted) and dialogue management (Miller, Hwang, Lee, Roberts,

Rudnicky, 2000). Led by Dr. Lee, her team of fifteen graduate students developed the BEE-SMART Semantic Web portal (BEE-SMART).

Dr. Poley, Executive Director of the American Distance Education Consortium, brings international experience, experience in adult distance education and existing collaborations with Native American tribal colleges. Dr. Poley's expertise is especially relevant in the area of development of a joint educational agenda. Dr. Poley has extensive experience in making joint teaching projects work over long distances, even when there are mismatches in technological implementation. Dr. Poley's American Internet Satellite Extension Project has been successful in bringing the Internet and Internet2 to remote and rural areas of the United States for purposes of improving research and education.

Dr. Soudi represents a multidisciplinary team from Morocco that is extensively examining knowledge-based machine translation methodology. That team includes expertise in computational and applied linguistics, computer science and engineering. Dr. Soudi has collaborative links with researchers in Germany and Italy, in addition to the United States.

Additional collaborators for the project have been identified at Kansas State University, University of Missouri at Columbia, Haskell Indian Nations University and Massachusetts Institute of Technology. In addition, the PI and Soudi will attempt to involve Soudi's collaborators Carnegie-Melon University.

It is the long-term intention to extend the collaboration beyond the US and Morocco, to other international teams with complimentary skills and/or common interests.

#### The Great Plains Network Consortium

The Great Plains Network (GPN) is a consortium of universities in the seven states of Arkansas, Kansas, Missouri, Nebraska, Oklahoma, North Dakota and South Dakota (<http://www.greatplains.net>, <http://research.greatplains.net>). Initially, the consortium was funded by the National Science Foundation to provide advanced networking services to member institutions. In the past three years, the consortium has become a member organization whose mission has shifted to working towards the establishment of large-scale joint projects, which are of interest to significant portions of the higher education membership in the region. The activities proposed, here, fit both the mission of GPN and the PI's major activities to network with faculty and other researchers at GPN member institutions to advance these new initiatives.

The GPN membership represents a diverse pool of talent in the areas of computer science, linguistics, psychology, cognitive science, education and special education, as well as augmentative systems and rehabilitation.

GPN membership has particular success in co-developing and team-teaching graduate courses, remotely, via the Internet. Kansas State University organized the first graduate course, taught simultaneously by faculty at University of Oregon, University of Nebraska and Kansas State University via Internet2.

### Languages of Interest

The initial languages of interest are English, Arabic and Berber. The main languages used in Morocco are Arabic and Berber.

Berber. According to statistics (atlapedia.com), Berber is the mother tongue of 40% of the people of Morocco. Linguistically, Berber belongs to the Afro-Asiatic group and has many dialects. The three main dialects used in Morocco are Tachelhit, Tamazight and Tarifit.

- Tachelhit is spoken in southwest Morocco
- Tamazight is spoken in the Middle Atlas
- Tarifit is spoken in the Rif area of northern Morocco.

Until 2001 Berber was not officially recognized in Morocco. All Moroccan school children are taught in Arabic, including children who speak Berber until reaching the classroom.

Arabic. The official language of Morocco is Arabic. Two forms of Arabic are used in Morocco: Colloquial Moroccan Arabic and Modern Standard Arabic. *Modern Standard Arabic (MSA)* refers to the variety of Arabic used in books, newspapers and media. Colloquial Moroccan Arabic (CMA) is the spoken form of MSA. One form of CMA is that used in educational and government settings. A second form of CMA (Median Moroccan Arabic, Youssi, 1986) is subject to regional differences; the lexicon of the second form derives not only from written MSA, but also from French and, in Northern Morocco, from Spanish.

### Broader Impacts

Impacts of the work to be done include

- Cultural: Assisting to safeguard linguistic and cultural diversity in the information society of tomorrow by strengthening the position of Arabic and other local dialects in linguistics and language technology;
- Social-Political: Promoting cooperation between the United States and (North) Africa for the purpose of developing basic components for the multilingual information society;
- Economic: Easing the entrance requirements for United States companies into the Moroccan market and vice versa by providing the basis for automatic translation tools which can be used to translate documentation and marketing material; and,
- Educational: Providing a platform for educational exchange and distance education leading to workforce development and increased

US-Morocco: Building a Human Language Technology Research and Education Agenda

Monaco, Soudi, Poley, Lee

economic opportunity for citizens as well as interdisciplinary, cross-cultural training of Human Language Technology graduate students.