Current Projects in Languages of Military Interest at the Defense Language Institute

Michael Emonts

Defense Language Institute Foreign Language Center
1759 Lewis Rd., Suite 245
Monterey, California, 93944-5006
Michael.Emonts@monterey.army.mil

Abstract

Recent US military events have produced a desperate need for language data in languages which have previously been given very little attention. Data resources for languages of military interest are often unavailable or insufficient to adequately develop new technologies. Realizing this need, the Defense Language Institute Foreign Language Center, has created a language data distribution center, which focuses on collecting and distributing data in languages of military interest. This paper outlines current projects and resources available at the Defense Language Institute (DLI).

1. Introduction

The language needs of the US military are constantly changing, since the Cold War model used in the past no longer applies in modern asymmetrical warfare. The interest has shifted from traditional Cold War languages, to low-density languages such as Pashto, Urdu, or Dari.

In an effort to better support the language needs of the Department of Defense, DLI has been increasing its involvement in language technologies. This paper outlines current language technology efforts and potential resources at DLI.

2. The Defense Language Institute

DLI, which is located in Monterey, California, is home to the world's largest foreign language instruction facility, run and operated by the United States Army. DLI provides highly trained linguists to all branches of the United States military, and is instrumental in the military's ability to provide national security.

Each year, DLI produces over 2000 graduates, with each linguist receiving 6-18 months of instruction depending on language difficulty. DLI also provides linguistic training and support for over 25,000 military linguists in the field through virtual training sessions and mobile training teams.

DLI is not only known for its ability to produce capable linguists, but is also well-known for its linguistic training and testing methodology. However, DLI is not only committed to providing linguistic training to the military, but aims to support all foreign language initiatives which help the Department of Defense meet its language needs. For this reason, DLI is becoming increasingly involved in language technology matters.

3. DLI Resources

3.1. Corpora

Since 1963, DLI has been involved in a wide variety of language projects, many of which may be of use to the language technology community. Previous language translation projects, for example, have generated large amounts of parallel texts which may be very beneficial to machine translation developers. Dictionaries, curriculum material, and linguistic studies may also be of use to NLP researchers, and are currently being collected and organized.

3.2. Native speakers

DLI employs over 800 native speakers in languages of military interest (Figure 1). Note that the languages listed in Figure 1 should not be viewed as a comprehensive list of languages for which the military has interest. In fact, many languages, which are indeed of high interest to the military are scarcely represented (or not all), due to the difficulty of finding qualified teachers of these languages. In general, DLI reacts to military requirements and finds faculty when needed.

Arabic	219	Italian	4
Korean	162	Albanian	3
Russian	122	Czech	3
Chinese	88	Greek	3
Persian/Farsi	46	Portuguese	3
Spanish	44	Uzbek	3
Serbo-Croatian	25	Armenian	2
French	12	Hindi	2
Hebrew	12	Tagalog	2
Dari	8	Urdu	2
German	8	Chechen	1
Thai	8	Georgian	1
Japanese	7	Ilocano	1
Pashto	5	Indonesian	1
Turkish	5	Javanese	1
Vietnamese	5		

Figure 1: DLI's Proficiency-rated Bilingual Speakers by Language *(as of October 2003)*

Many DLI instructors are highly educated language experts who, in addition to being available for data collection, also offer linguistic expertise.

When dealing in highly-sensitive military matters, such as force protection, speakers from some areas of the world often only reluctantly participate in US government-sponsored projects. DLI employees, however, are not only willing to cooperate on military matters, but are also generally well-educated in the military domain.

Furthermore, since an extensive background check is a necessary prerequisite for working at DLI, data collected from DLI employees can generally be regarded as more reliable. This is a significant issue in the government, since the loyalties of foreign nationals are not always clear, and ulterior motives may exist when participating in military-sponsored projects.

3.3. Defense Language Proficiency Test (DLPT)

As part of the interviewing/screening process, DLI instructors undergo extensive language testing to grade their linguistic abilities — not only in their native language, but also in English and any other languages they may know. Students at DLI are frequently tested, both to monitor their progress, and to check if they meet graduation requirements.

The DLPT is a test developed at DLI for the testing of language proficiency. For each test subject, the DLPT produces a score from zero to five for each of the four language skills (reading, writing, listening, and speaking) on the widely-used Interagency Language Roundtable (ILR) rating scale (Child, 1998). Detailed descriptions for each language level have been rigidly defined for each language skill, and can be viewed at http://www.monterey.army.mil/atfl/daa/skill.htm. Figure 2 shows a general description for each language level.

0	No Proficiency
0+	Memorized Proficiency
1	Elementary Proficiency
1+	Elementary Proficiency Plus
2	Limited Working Proficiency
2+	Limited Working Proficiency Plus
3	General Professional Proficiency
3+	General Professional Proficiency Plus
4	Advanced Professional Proficiency
4+	Advanced Professional Proficiency Plus
5	Functionally Native Proficiency

Figure 2: ILR Rating Scale

The DLPT is of great use to the language technology community, because it ensures that all data collected at DLI is not only from highly-proficient speakers, but from speakers with known competency ratings. This is quite different from many current data collections, where minimal effort is applied to verify language competency.

DLPT scores are even more important for languages of military interest, since many of these languages are often without any standardized writing system, and come in many dialect variations. Therefore, data collected from a native Pashto speaker, for example, might be characterized by long pauses and frequent mispronunciations because of the speaker's inability to read the Pashto-scripted prompts. In this case, it would have been extremely useful to know the speaker's reading ability beforehand.

Having access to a speaker's DLPT scores for all four language abilities allows one to predetermine the qualification criteria for the speakers, and reduce the chance of undesirable data entering the corpus — which might go unnoticed otherwise.

4. Language Data Distribution Center

Because of the high demand for data in military languages, and the lack of a centralized location to house and distribute language data, the DLI Language Data Distribution Center (DLI-LDDC) was created.

The DLI-LDDC is committed to becoming the US government's central repository for language data, in direct support of government-sponsored research. Language corpora at DLI-LDDC are distributed at no cost, but are only available to government entities and organizations with government contracts who qualify for access.

Even though much of the data used in military applications is either unclassified or declassified, military language data can often be of sensitive content, not intended for the public. For this reason, the DLI-LDDC has been created with high security in mind. Data is housed in a secure-access, military-controlled facility, and is protected by a backup generator, a temperature control system, antistatic floors, and a dry fire suppression system.

Data transmission through NIPRNet (non-classified internet protocol router network) and SIPRNet (secret internet protocol router network) facilitates distribution to government agencies. For further convenience, a 1-GB uplink to the Internet has been established, as well as a user-friendly, dynamically scalable Internet delivery system.

5. Current Projects

5.1. DARPA CAST (formerly Babylon)

Military forces need to verbally communicate with the local populace, enemy POWs, displaced civilians, refugees, and non-English speaking coalition forces. In the absence of a proficient foreign language specialist, speech translation devices can help meet this need. Speech translation systems enhance the soldiers' ability to gather vital intelligence, develop working relationships with the locals, as well as perform their regular duties which involve communication with non-English speakers.

The goal of the Compact Aids for Speech Translation (CAST) program is to develop rapid, two-way, natural language speech translation interfaces and platforms for the warfighter engaged in force protection, refugee processing, and medical triage. CAST focuses on overcoming the technical challenges which currently pose limits on multilingual translation technology. Current languages being developed are Farsi, Mandarin, Pashto, and Thai.

DLI's primary role in the CAST program is to translate military force protection and medical triage dialogs for the development teams. DLI was also charged with the task of establishing a data repository (i.e. the DLI-LDDC) to house the data -- which may then be distributed for use in future projects.

To date, 630,000 words have been translated into Mandarin; 340,000 words have been translated into Farsi, and a Thai translation project is set to begin in the near future.

5.2 Machine Translation Evaluation

DLI is currently sponsoring a collaboration with MIT Lincoln Laboratory on a research project to examine the relationship between automated measures of machine translation quality (Doddington, 2003) and the ILR scale of language proficiency (Child, 1998), which has been a government standard for language training and assessment for the past several decades.

Difficulty level of the input text is typically not addressed in many current automated MT scoring techniques. Input text difficulty is being analyzed with respect to MT performance and scoring methods. The official NIST MT Evaluation scoring package based on the IBM BLEU scoring tool (Papineni et al., 2002) is used as a test for this study.

For this research, DLI also provided the SPARK (Spanish, Persian, Arabic, Russian, and Korean) corpus, which is a small corpus of text documents at each of seven difficulty levels (1, 1+, 2, 2+, 3, 3+, and 4), across a range of topical domains. Each text in the SPARK corpus is accompanied by an English source text and a commentary on the difficulty level. A variety of experiments were performed to examine the relationship between difficulty level of input text and machine translation performance (Clifford et al. 2004).

5.3. Bilingual Speech Corpus Collection

BAE Systems partnered with DLI for a 3-month data collection effort to develop a multi-language speech corpus to be used for bilingual speech recognition and voice modeling. DLI provided 238 bilingual speakers for this project (15 Spanish-speakers, 115 Arabic-speakers, 108 Korean-speakers).

The data collected was telephone-based conversational style recordings, with each speaker providing two recordings with approximately a month separating the two sessions.

5.4. The Mixer Corpus

The Linguistic Data Consortium (LDC) recently completed the Mixer corpus (Cieri et al., 2004), which is a large collection of multilingual, multichannel speech, involving over 600 speakers participating in up to 25 calls each, often involving unique telephone handsets. Data was collected in English, Arabic, Russian, Mandarin, and Spanish.

DLI helped by providing native Arabic-, Chinese-, and Russian-speakers. 24 DLI employees participated in this data collection effort by making a total of 387 phone calls.

6. Conclusion

DLI has many resources of interest to the language technology community. Aside from language corpora, and proficiency-rated native speakers in language of military interest, DLI also has the DLPT and ILR scale to add reliability to machine translation evaluations. With the newly created language data and distribution center, DLI has enhanced its ability to collect and distribute language data.

7. References

Child, James R. (1998). Language Skill Levels, Textual Modes, and the Rating Process, Foreign Language Annals, 31, NO. 3.

Cieri, Christopher, David Miller, Kevin Walker. (2004). The Mixer Corpus of Multilingual, Multichannel Speaker Recognition Data. In Proceedings of LREC 2004. http://www.ldc.upenn.edu/Mixer/

Clifford, Ray, Neil Granoien, Douglas Jones, Wade Shen, Clifford Weinstein. (2004). The Effect of Text Difficulty on Machine Translation Performance – A Pilot Study with ILR-Rated Texts in Spanish, Farsi, Arabic, Russian, and Korean. In Proceedings of LREC 2004.

Doddington, George. (2004). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. National Institute of Standards and Technology Machine Translation Evaluation Web Site, February 2003. http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf

Papineni, Kishore, Salim Roukos, Todd Ward, Wei-Jing Zhu. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In Proceedings of ACL 2002. (pp. 311--318).