Language Learning Data: Online Confusion

Flo Reeder & Lisa Harper
The MITRE Corporation
1820 Dolley Madison Blvd.
McLean, VA 22102
703-883-7156 (voice)
703-883-1379 (fax)
[freeder, lisah@mitre.org]

Abstract

Corpora, particularly resources available through the World Wide Web (WWW), are a potential gold mine of teaching materials, especially for Computer-Assisted Language Learning (CALL). These resources can be static, such as newspaper repositories; partially interactive, such as list servers and mailing lists; or completely interactive, such as chat rooms. While this vast data resource has the great promise of diverse and interesting materials for the language learner, it also suffers from some basic technological problems that can hinder language-teaching efforts. Most of these stem from the fact that much computing technology is not internationalized. The ASCII bias of many systems means that language representation is confusing and difficult. For instance, the discussion list French Talk, designed as practice for language learners, displays three separate strategies for handling accented characters: ISO Standard 8859-1 encoding, ASCII transliteration and accent omission. This is a pervasive problem in many domains of Natural Language Processing (NLP), yet it can be claimed that these errors place an additional burden on the language learner. We argue, therefore, that teachers and administrators should be aware of these issues and that resource providers need to supply tools for automatically cleaning up text. This paper presents the kinds of problems we have found in some sample resources and makes some preliminary recommendations for avoiding them.

Keywords: Computer-Assisted Language Learning, Corpus Analysis, Text Correction.

1 Introduction

Online resources are plentiful and increasing. This presents a boon for teachers of less commonly taught languages. Types of readily available resources include electronic dictionaries, electronic corpora and interactive media. Corpora consist of newspaper repositories, online journals and magazines and literary archives. Interactive opportunities are growing by leaps and bounds with the growing presence of Internet Relay Chat (IRC), "I seek you" (ICQ), web-enabled Java chat-based environments, online distance learning, virtual universities, discussion lists and news groups. The inaugural meeting of WorldCall (WorldCall 1998) highlighted over 20 projects utilizing interactive web technology for language learning.

2 Previous Work

With her evaluation of Machine Translation (MT) systems, Flanagan (1994, 1996) describes reasons for errors in translation and discusses some of these errors in terms of real-world

translation expectations. The evaluation parameters of these MT systems were examined in light of the translation output and the type of error generated by these various translation engines. These include spelling errors, words not translated, incorrect accenting, incorrect capitalization as well as grammatical and semantic errors. These types of error reflect the quality of input into the translation process. Flanagan argues for pre-editing tools such as spelling checkers and grammar checkers.

Kukich (1992a, 1992b) also presents many sources of error-filled input for language processing systems. These range from simple typing errors to human cognitive errors. While she addresses only English-language errors, multi-lingual environments add another dimension towards complicating error analysis. In addition to traditional insertion, deletion, transposition errors, multi-lingual data can contain transliteration, transcription and code set representation errors. Kukich's work, in combination with Flanagan's analysis, motivated us to analyze and evaluate the problem of "not-translated" words in real-world translations.

Due to the explosive availability of "real-world" foreign language materials on the internet today, all of these types of error-filled input that create challenges for NT and NLP systems also present problems for language learners.

3 Approach

For this study, we collected a set of e-mail documents from discussion lists, World Wide Web (WWW) documents, IRC logs, electronic news articles and general texts. We then attempted to identify the encoding or transliteration scheme and locate fonts (if necessary) to view these documents properly. A significant challenge was the necessity to configure multiple application on individual computers in order to read and write text in various languages. For example, in order to read and write in Thai on Microsoft Windows 95, we had to configure a mail application and WWW browser, upload an initialization file for an IRC browser, find software to toggle keyboard mappings, and adapt system configuration files and fonts for telnet.

4 Results & Analysis

The types of problems prevalent in online resources can be categorized into two major categories: *availability of data* and *quality of data*.

In the first category, we found problems such as insufficient bandwidth supporting language-learning interaction. This is especially true of audio and video resources such as sound clips and video streams. Frame speed can also be a source of distraction in video teleconferencing (VTC). While these technologies are rapidly becoming more sophisticated and sensitive to fluctuating bandwidth conditions (e.g., RealAudio adaptive streaming techniques), choppiness and slow frame rates are still causes of common complaints. Additionally, we found copyright restrictions and platform availability encumbered the use of many of these tools.

Of more interest to us, as NLP researchers, is the issue of resource quality. Beginning language learners require very good quality materials. In the absence of good quality materials, language learners become distracted and frustrated by peripheral issues of having to convert materials into usable form. For instance, while gathering foreign language materials on the web, we frequently found inconsistent text representations. Whereas the computing world is set up to handle English and most Western European languages without great difficulty, transmitting and computing in languages such as Chinese, Japanese, Thai and Russian present great difficulties. Even a Western European language such as French or German causes problems. In data gathered from FrenchTalk (a French language discussion list for language learners), we found four commonly used representations for French words: ISO-8859-1, two different schemes for

ASCII transliteration, and plain ASCII with no complicated representational problems. For encodings (Code Page 1251, KOI-8, ISO-5 and transliterations and transcriptions (as many as particular difficulty when one looks at (see Figure 1). In fact, considerable research devoted to accurately deciphering transliteration of transliteration schemes also presents languages associated with multiple encoding or forced to find a way to identify the code set or

*** Now talking in #Thailand
<yaiBow> waddee ka kiat
<^^^IRENE> ËÇÑ´´Õ¾Õèà¡ÕÂÃµÔ
<^^^IRENE> ÇÑ¹¹Õé¨Ðä»äË¹ÍÕ¡ÍêÐà»ÅèÒ¤Ð
<kiat> ¤×¹¹ÕéÇÑ¹ËÂØ´..áÅéÇ
<^^^IRENE> äÁèä»äË¹
<yaiBow> hi gromitbear
<^^^IRENE> ¾Õèâº¢Ò
<^^^IRENE> ¾Õèà¡ÕÂÃµÔã¨´Õ
*** ^voy^ (7000@hlr-13-125.tm.net.my) has joined #thailand
<^^^IRENE> ¾Õèà¡ÕÂÃµÔ ¾Õèâº ¹èÒÃÑ¡
<kiat> yaiBow waddee krab'
* fgel ã¨ÃéÒÂ µÒàºÅÍ
<gromitbear> hi. how can you read in thai?
-ChanServ:#Thailand-
kiat!PaNZeR@161.246.11.204 opped ^^^IRENE
*** ChanServ sets mode: +o ^^^IRENE

Figure 1: Chat on a Thai IRC Channel

The language learner's problem does not end after she has identified the encoding or transliteration scheme. Encodings are rendered on a display device by the mapping an encoding to a specific kind of font. This generally means that fonts must be downloaded and set up on the computer. Even if the code set is properly identified, it may not automatically readable by the computer as illustrated in Figures 2 and 3. Additionally, computing processes introduce errors into the mix as shown in table 1. Once text is readable, there is still the issue of production. Learners must learn to negotiate a new set of keyboard mappings if they wish to type text in a word processor.

Nj& burim i LDK-s& nga Llausha tha se ka njoftime p&r shtim t& forcave serbe n& af&rsi t& Turi^ecit, n& hap&sir&n mes k&tij fshati e Llaush&s. Duke folur p&r pasojat e jet&s n&n rrethim t& plot& policor, me munges& ushqimi e ila^esh, ai tha se Adem Rrecaj, nj& 75- vje^ar nga ky fshat, q& ishte I s&mur& edhe m& par&, vdiq ngase nuk pati mund&si askush t'I ofroj& ndihm& mjek&sore.

Figure 2: Albanian Transliteration

Një burim i LDK-së nga Llausha tha se ka njoftime për shtim të forcave serbe në afërsi të Turiçecit, në hapësirën mes këtij fshati e Llaushës. Duke folur për pasojat e jetës nën rrethim të plotë policor, me mungesë ushqimi e ilaçesh, ai tha se Adem Rrecaj, një 75- vjeçar nga ky fshat, që ishte I sëmurë edhe më parë, vdiq ngase nuk pati mundësi askush t'I ofrojë ndihmë mjekësore.

Figure 3: Albanian Corrected Text

Error type	**Example**
Scanning error	écrire → 6crire
Transmission error	écrire → =E9crire
Misplaced control character error	This is ^Z the end.
E-mail address	freeder@123.mit.org
HTTP Markings	</end>
Stripped bits	écrire → hcrire

Table 1: Some Common Types of Translation Error in Multi-lingual Text

To determine how pervasive these quality problems are, we looked at one particular language resource: the on-line discussion list FrenchTalk. It is a resource designed specifically for language learners. We collected data from this list for about 6 months and were overwhelmed by the amount of material: we gathered more than 5MB of text. This source exhibited many of the problems we have discussed: absence of diacritic markings, mixed language documents, odd representations (such as the Eudora quoted-printable), misspellings and inappropriate or offensive topics of discussion. Figures 4 and 5 show extracts from FrenchTalk and exhibit a variety of error types.

<frenchtalk@list.cren.net>
Subject: L'age de la petite fille du capitaine

À (At) 5:38 –0400 15/07/97, JoeCool@aol.com écrivait (wrote) :
>Oh non, elle n'est pas aussi agee que ca!
>
C'est vrai... c'est juste pour que Reynald ne bave pas derriere son ecran. ;-)

Michele

Read you soon on the Moon

Figure 4: Extract from FrenchTalk

These texts include examples of mixed language and the inclusion and absence of diacritics. Although the content in this note is sufficient for most first and second-year learners to understand the meaning of the machine "not translatable" words, there are circumstances when the absence of diacritics renders a sentence completely ambiguous despite context. For example, the Slovenian sentence:

Problem je resen

means alternatively *The problem is a serious one* or *The problem is solved*, if the accent above the *s* is omitted.

In the example below, the first writer (left) sent a message using ISO-8859-1. The message may have been readable as such by the person who responded with the text reply to the right, but the reply was in ASCII formatting and all of the diacritics from the original text were lost. Language learners require more context in order to understand words that they are less familiar with. Diacritics can provide valuable context for disambiguating words.

>Save Our Screen...>Je sais que l'âge d'or des screensavers est fini depuis longtemps mais je >cherche un screensaver, petit, sympa, sobre, élégant passque ma bécane est
>dans le salon et que voir "mon bureau" after hours me D' !

>Save Our Screen...>Je sais que l'age d'or des screensavers est fini depuis longtemps mais je>cherche un screensaver, petit, sympa, sobre, elegant passque ma becane est
>dans le salon et que voir "mon bureau" after hours me D' !

Figure 5: Two extracts from FrenchTalk. The one the left was typed in ISO-8859-1. The response on the right includes a copy of the original text but was sent in an ASCII reply.

IRC demonstrates different sorts of problems due to the nature of real-time text-based communication. Chat allows students to become more interactive at an earlier stage in their learning, but brings with it yet another application to configure and new types of errors for students to encounter. The example from a German chat room below exemplifies multiple languages, the absence of diacritics, and the absence of upper casing on German verbs. Case is contextual information relevant for students learning to distinguish nouns.

*** GNZ (wilma@200.36.51.177) has joined #germany
<philipp> hab ich das topic gesetzt? ich hab dich nur nach deiner meinung gefragt
<vonLUNEN> VOUS PARLEZ FRANCAIS? HMM?? :)
<philipp> achso das mit dem bot
<philipp> ja vergiss das

<vonLUNEN> il est tres difficile comprende votre langue
<MrAbd> Hi
<Ken_cafe> ich meine ja nur
<vonLUNEN> Mr Abd.. ya Hala
*** jaw0 (~x5algh@141.35.2.80) has joined #germany

Figure 6: Chat on a German IRC channel

5 Conclusion & Future Work

It could be argued that some of these types of problems would disappear with the advent of tools and technologies such as Unicode, improved operating system support and other advances. While industry movements towards Unicode and ISO 10646 are a step towards eliminating these problems (Adams, 1993), existing systems will leave a legacy of errors for new generations. Additionally, the variability between these two standards has not sufficiently converged to allow for quick and easy language determination.

Currently, language teachers and learners must invest a considerable amount of effort finding and installing fonts and keyboard mappings and configuring software settings for browsers, mailers and chat software on a case-by-case basis. Ideally, we would like to allow students to sit down at any terminal, open an application, and be able to view and create text materials in a more standardized way.

What can be done, then, to help the language learner navigate through a sea of electronic data? First, the development and introduction of automated processing tools for areas such as code set identification, code set conversion, diacritic correction, spelling correction and grammar correction is a necessity as language learners use available online tools. Second, access to dictionaries and corpora needs to be standardized and made as accessible as possible. Third, we must actively lobby for policies allowing the inexpensive use of materials for educational purposes. Finally, teachers, institutional media administrators and courseware developers need to be made aware of the pervasive problem of language representation. They must become more informed about strategies aimed towards reducing the complexity of foreign language materials presented to students.

References:

(Adams 93) G. Adams, *Internationalization and character set standards*, Standard View, 1(1), 31 – 39, 1992.

(Bech 97) A. Bech, *MT from an everyday user's point of view*, MT Summit, pp. 98-105, 1997.

(Flanagan 94) M. Flanagan, *Error classification for MT evaluation, Technology Partnerships for Crossing the Language Barrier*, Proceedings of the First Conference of the Association for Machine Translation in the Americas, Columbia, MD, 1994.

(Flanagan 96) M. Flanagan, *Two years online: experiences, challenges and trends*, Expanding MT Horizons: Proceedings of the Second Conference of the Association for Machine Translation in the Americas, (pp. 192-197), 1996.

(Knight & Graehl, 97) K. Knight, J. Graehl, *Machine transliteration*, Proceedings of the 35th Annual meeting of the Association of Computational Linguistics, 1997.

(Kukich 92a) K. Kukich, *Techniques for automatically correcting words in text.* ACM Computing Surveys, Vol. 24, No. 4, 1992.

(Kukich 92b) K. Kukich, *Spelling correction for the telecommunications network for the deaf*, Communications of the ACM, Vol. 35, no. 5, pp. 80-90, 1992.

(Kumhyr et al. 94) D. Kumhyr, C. Merrill, K. Spalink, *Internationalization and translatability*, Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas, Columbia, MD, 1994.

(WorldCALL 98) WorldCALL98: Call to Creativity, University of Melbourne.

(Yarowsky 1994) D. Yarowsky, *Decision lists for lexical ambiguity resolution: application to accent restoration in Spanish and French*, Proceedings of the 32nd Annual Meeting of the Association of Computational Linguistics, 1994.