Message from the Chairs

ILRIG had a fantastic program and business meeting at the ASIL 2014 Annual Meeting. Our special thanks go to Marylin Raisch and her wonderful speakers: Ana Ayala from the O'Neill Institute (Georgetown Law), Prof. Jeffrey Ritter from Georgetown Law, and Dr. Alejandro Ponce from the World Justice Project.

Report on the 2014 Business Meeting:

The business meeting was short and sweet. Gabriela Femenia, our ILRIG Treasurer, gave a report on our budget. Marylin Raisch, our ILRIG Secretary, gave a brief description of the award that we would like to begin offering (the Jus Gentium Award, described below). She requested input from the members on ideas about the award's name, the criteria for awarding it, and the choosing of winners.

ILRIG also requested that anyone interested in co-editing the newsletter should contact any ILRIG officers. Please notify one of the co-Chairs for details. Prior editions of *The Informer* are available through the ILRIG website. *The Informer* was cited as a model ASIL interest group newsletter at the ASIL interest group Chair breakfast in 2012.

We had a vote on the proposed amendment to the ILRIG bylaws requiring that at least one of the co-chairs be an FCIL law librarian. The vote was: 8 in favor, 0 against, 2 abstentions. The proposed amendment passed. We then asked for questions, comments, or announcements from the floor.

An announcement was made congratulating Lyonette Louis-Jacques on the publication of *International Law Legal Research,* of which she is a co-author.

Update on the Jus Gentium Research Award

Since the last Annual Meeting, IL-RIG has proposed the creation of the Jus Gentium Research Award for the recognition of important contributions in the area of providing and enhancing open access to legal information resources in inter-

NEWS & EVENTS
 ILRIG Business Meeting Thursday April 9, 2:45-4:15 p.m., Glacier room @ Hyatt Regency Capital Hill

**INTEREST GROUP OFFICERS
CO-CHAIRS:**

 Jootaek ("Juice") Lee
 Wanita Scroggs

SECRETARY:

 Marylin Raisch

TREASURER:

 Gabriela Femenia

NEWSLETTER EDITORS:

 Don Ford
 Jootaek ("Juice") Lee, Guest Editor

national law. The Jus Gentium Award has subsequently been approved by the ASIL leadership. Accordingly, ILRIG created an award selection committee consisting of ILRIG members from among academic institutions, law firms, IGOs, NGOs, and international and domestic courts. This year, Lyonette Louis-Jaques (Chair; University of Chicago School of Law), Freddy Sourgens (InvestmentClaims), Paulina Starski (Max Planck Institute for Comparative Public Law and International Law), and Muruga Perumal Ramaswamy (University of Macau) are serving on the committee. We sincerely extend our gratitude to the committee members. The award will be granted to the creators—either individuals or institutions—of non-commercial, on-line databases freely available to the international law community as well as to the public at large, enhancing both scholarship and open access to legal information.   The selection committee will choose one recipient each year who will then receive a commemorative plaque of recognition from ILRIG during the Annual Meeting.

Update on the Research Liaison Program

ILRIG will continue to offer research services through the Research Liaison Program for the speakers and moderators at the ASIL Annual Meeting. Services will be available prior to and during the meeting. However, the physical research kiosk at the Annual Meeting will be discontinued in 2015.

We are also pleased to remind ILRIG

members that the Kiosk in 2014 was sponsored by HeinOnline, which graciously provided complimentary access to the Hein databases. We are extremely grateful to HeinOnline for their generosity. These materials significantly enhanced the services that the ILRIG Research Liaison Program was able to provide.

This year we saw the arrival of the ILRIG webpage on ASIL's new website, and we are very excited about this wonderful way for members to keep in touch.

ILRIG continues to submit program proposals for the ASIL Annual Meeting, and continues to sponsor/co-sponsor meetings and webinars with other professional organizations.

A Special Message from ILRIG Co-Chair Wanita Scroggs:

The 2015 Annual Meeting marks the end of a busy three year term for co-chair Jootaek "Juice" Lee.   ILRIG thanks Juice for his service during these years of tremendous growth and change for our interest group. And we will wish our new co-chair (to be elected) continued success and a warm welcome to the ILRIG leadership.

We are looking forward to seeing you during April 2015! Thank you!

Co-chairs:

Jootaek ("Juice") Lee
joo.lee@neu.edu

Wanita Scroggs
wscroggs@law.stetson.edu

"We hope that Part II of our article will dispel doubts about Machine Translation (MT) and Computer Assisted Translation (CAT)."

A Primer on Machine Translation (MT)

Part II: The Technical Aspects of Machine Translation (MT), with a Select Bibliography of Law-Related MT Resources

By Don Ford and Matthew Gran[1]

Abstract: The **Informer** continues with Part II of "A Primer on Machine Translation" (Part I is in the Winter 2013 *Informer* at http://www.asil.org/sites/default/files/documents/ Informer_Winter_2013_Issue.pdf). In this issue, co-author Matthew Gran explains the technical aspects of Machine Translation (MT), including rules-based MT and statistical MT. Co-author Don Ford then gives examples of the use of MT in everyday work usage: the translation of foreign language bibliographic records.

_____

Technical Aspects of Computer Assisted Translation

Perhaps you have used a machine translation ("MT") program like Google Translate or Babel Fish to translate a legal text.[2] Some outputs were good (i.e., close enough) translations that accurately depicted the meaning of the inputted text. Other times, the outputs were downright strange: garbled text that made no sense. You might even have had good and bad results for the same inputted text.[3] Understandably, it is easy to ascribe these mixed results to deficient technology. It is also easy to vow to never use MT programs again. However, MT and other Computer Assisted Translation ("CAT") systems continue to improve.[4] We hope that Part II of our article will dispel doubts about MT and CAT by providing a methodology for approaching MT and CAT programs.[5]

_____

[1] Don Ford is the FCIL Librarian at the University of Iowa College of Law Library. Matthew Gran is judicial law clerk for the Honorable Eileen Mary Brewer, Judge of the Illinois Circuit Court of Cook County (Law Division). The authors wish to thank Harold Somers for both his helpful and critical remarks and suggestions during the writing of this article. Professor Somers is Professor of Language Engineering (Emeritus) at the University of Manchester, United Kingdom. In addition, the authors thank Roy Sturgeon for his help with basic Chinese Pinyin. Mr. Sturgeon is Foreign, Comparative, and International Law/Reference Librarian at Tulane University Law Library.

[2] MT is the general term encompassing computer programs that automatically translate from one language to another. *See* Ross Smith, *Machine Translation: Potential for Progress*, 17 ENGLISH TODAY 38, 38 (2001). MT has several subclasses, which will be discussed later in this section.

[3] *See* Nicola Cancedda et al., *A Statistical Machine Translation Primer, in* LEARNING MACHINE TRANSLATION 1 (Cyril Goutte ed., 2009).

[4] *See* Alexandra Robbins, *Mining for Meaning: A Renaissance for Machine Translation*, PC MAGAZINE 25 (Dec. 30, 2003, 12:00 a.m. EST) , http://www.pcmag.com/article2/0,2817,1401163,00.asp; Lawrence Williams, *Web Based Machine Translation as a Tool for Promoting Electronic Literacy and Language Awareness*, 39 FOREIGN LANGUAGE ANNALS 565, 565 (2006). CAT is the linguistic and computer programming term encompassing translation software, including MT programs. Sometimes, though, MT is distinguished from CAT because MT automatically translates from one language to the next. Still, MT programs are typically used to assist with translation because translators and users often review and edit MT outputs. Smith, *supra* note 2, *at* 38, 40.

[5] *See* Robbins, *supra* note 4; Williams, *supra* note 4, at 565.

There are different types of CAT, ranging from machine translation ("MT") systems, translation memory ("TM")[6], concordances[7], and dictionaries.[8] It is important to question a given CAT program's quality and usefulness because these factors differ by project. Output quality can depend on the type of translation being done, the expectations of the program's user, and/or the eventual user of the translated end product. Additionally, the accuracy of a translation often will vary more, depending on the scope of the translation, ranging from discrete words and phrases, to passages, to integrated documents.[9]

Legal texts are particularly difficult to translate.[10] Law is a complex domain that is highly specialized. A word or legal phrase's meaning varies by context, and may require a certain level of legal expertise to understand its meaning.[11] And, legal concepts and terminology change over time.[12] While CAT cannot effectively translate *entire* legal texts, it can *assist* in legal translation.[13] In order to employ CAT effectively, users should evaluate their own language proficiencies, command of legal concepts,[14] and understanding of different CAT systems.[15] This self-evaluation is critical when deciding whether to use CAT.[16]

The following figure (Figure 1) shows the benefits CAT users get from evaluating their foreign language reading proficiency, with suggestions regarding the use of CAT.

_____

[6] A translation memory program stores previous translations and enables translators to reuse prior translations. TM can scan a document and identify and translate sentences or phrases that were previously translated and stored in the program's database. Smith, *supra* note 2, at 38.

[7] "Concordancers" are programs that can construct a concordance for a set of translated documents for a language pair. Concordancers are particularly useful because they provide prior translations, giving examples of the different ways a word or phrase was translated. Marilyn Domas White et al., *Beyond Dictionaries: Understanding Information Behavior of Professional Translators*, 64 JOURNAL OF DOCUMENTATION 576, 595-96 (2008). Legal translators can utilize Linguee.com (http://www.linguee.com) as a concordancer because the site incorporates parallel translations of legal and government texts from sources such as the European Union and the United Nations.

[8] Dictionaries range from bilingual to multilingual. *See id.* at 595. For example, the European Union's InterActive Terminology for Europe is an excellent multilingual dictionary that can be used to translate legal words and phrases. The dictionary has a dropdown box that allows a user to limit the domain to law. *See* INTERACTIVE TERMINOLOGY OF EUROPE, http://iate.europa.eu (providing a multilingual dictionary between 25 European languages).

[9] Harold Somers, *The Use of Machine Translation by Law Librarians—A Reply to Yates*, 99 LAW LIBR. J. 611, 612 (2007).

[10] Chunyu Kit & Tak Ming Wong, *Comparative Evaluation of Online Machine Translation Systems with Legal Texts*, 100 LAW LIBR. J. 299, 305 (2008).

[11] Somers, *supra* note 9, at 613-614.

[12] Pierre Mazzega et al., *A Complex-System Approach: Legal Knowledge, Ontology, Information and Networks, in* APPROACHES TO LEGAL ONTOLOGIES 117 (G. Sartor et al. eds., 2011).

[13] Somers, *supra* note 9, at 619; Kit & Wong, *supra* note 10, at 320-321; Roy Balleste, *Could You Translate the Site For Me?: Rating the Usefulness of Translation Software Programs for Law Library Web Site*, 5 AALL SPECTRUM 4, 4 (Jul. 2001).

[14] Kit & Wong, *supra* note 10, at 305.

[15] *Id.* at 321.

[16] *Id.*

Proficiency ratings and measurements are taken, adapted, and condensed from the Inter-agency Language Roundtable's[17] contrasted reading proficiency scales.[18] Individuals with greater language proficiency can use CAT effectively to translate foreign legal texts because they are able to edit and correct inaccurate translations. Individuals with little proficiency, though, are less able to spot inaccuracies.

Language Proficiency and MT Effectiveness

No Knowledge   Elementary   Limited Working   General   Advanced   Native

Low Effectiveness   Highly Effective

Fig. 1

Language proficiency should be determined by the individual's ability to translate text from the source language to the target language.

1. No Knowledge or Proficiency: These users should not depend on CAT to translate legal texts. While CAT can provide access to otherwise inaccessible text, these users cannot easily correct inaccurate CAT outputs.

2. Limited Working Proficiency: These users possess enough proficiency to understand the language they are reading. These individuals can benefit from using CAT because they understand the language's grammar and syntax well enough to determine whether the program has applied the language rules correctly. Still, these users will likely miss mistranslated phrases or legal terms.

3. General Professional Proficiency: These users are able to sufficiently read most things but may miss some technical aspects or subtleties involved in a text.

4. Advanced Professional Proficiency: Foreign language skills are proficient enough to communicate and read for professional needs. Person is unlikely to misinterpret technical aspects or subtleties of a paper. Users can use CAT effectively, but may miss or overlook subject matter specific terms, like legal terms.

5. Functionally Native Proficiency: These users can benefit the most from using CAT because they can identify and correct incorrect outputs. Additionally, legally trained users will most likely know if a legal term was translated incorrectly, and are able to use a CAT program to verify whether a term or phrase was translated correctly.

---

[17] The InterAgency Language Roundtable (ILR) is a federal program that encourages cross-agency sharing and development of foreign language training and testing. *See About the IRL*, available at http://www.govtilr.org/IRL%20History.htm (last visited Mar. 9, 2015).
[18] InterAgency Roundtable, *ILR Reading Skill Scale*, available at http://www.govtilr.org/skills/ILRscale4.htm (last visited Mar. 9, 2015).

For example, a law librarian is researching an international legal subject. The research includes citations to several non-English language legal documents. If there is no translation available, the librarian may want to screen the document to see if a translation is warranted.[19] Here, a law librarian with the relevant language proficiency could use MT to quickly translate selected passages and the librarian would then decide whether a full, professional translation is warranted. The law librarian without the relevant language proficiency could use CAT; however, this individual would likely miss output errors.[20]

Users who understand the linguistic aspects of English and the foreign language in the above example are positioned to use CAT effectively. For these users, "close enough translations" are good enough, because users can manually correct inaccuracies.[21] As long as the corrections do not take more time than manual translation, CAT would increase the user's productivity.[22] Applying this methodology to legal texts is not easy, especially with MT. CAT programs are often not equipped to translate legal words and phrases because of several linguistic issues: lexical ambiguity,[23] syntactical level,[24] specialized language problems,[25] lexical gaps between languages,[26] differing national legal frameworks for similar legal issues, and terminology misalignment.[27]

Professionals with sufficient language skills and legal expertise can identify these issues, enabling them to correct inaccurate translations. However, these professionals still struggle with errors caused by CAT, specifically with MT programs, which are increasingly being used by professional translator services.[28] MT programs are not transparent and reasons for translation errors, especially ones the system is supposed to translate, remain elusive to these professionals.[29]

---

[19] Somers, *supra* note 9, at 618; Kit & Wong, *supra* note 10, at 321.

[20] Lawrence Williams, *Web Based Machine Translation as a Tool for Promoting Electronic Literacy and Language Awareness*, 39 FOREIGN LANGUAGE ANNALS 565 (2006).

[21] *Id.* at 575-576.

[22] Craig Morgan Teicher, *Prompting a New Way To Translate: Is a machine a viable way to translate some books?*, PUBLISHERS WEEKLY 16, 16 (Apr. 20, 2009).

[23] Somers, *supra* note 9, at 612 (describing how lexical ambiguity occurs with a word with "a variety of meanings").

[24] *Id.* (defining syntactic ambiguity as a series of words with multiple interpretations; additionally, Somers provided the frequently quoted syntactic ambiguity example of "Fruit flies like a banana").

[25] *See* Mazzega et al, *supra* note 12, at 127 (describing how the legal domain provides specialized meaning to terms. Here, the authors illustrate how a statistical relevancy based analysis of a legal oncology can associate a similar meaning to two terms, even though both words have a distinct legal meaning).

[26] Kit & Wong, *supra* note 10, at 304 (describing how a lexical gap is created between a linguistically diverse language pairing).

[27] Gianmaria Ajani et al., *Terminological and Ontological Analysis of European Directives: Multilinguism in Law, in* ELEVENTH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW 43, 44 (2007) (noting that different countries' legal frameworks can define a legal concept differently and also the use of a term and the translation can vary across different areas of law).

[28] Nikiforos Karamanis et al, *Translation Practice in the Workplace: Contextual Analysis and Implications for Machine Translation*, 25 MACHINE TRANSLATION 35, 45 (2011).

[29] *Id.* (discussing how a translator found that the MT program failed to translate certain strings within a passage for no apparent reason).

Further, MT has been criticized as not being able to overcome the above-mentioned linguistic challenges.[30] Still, MT is a computer application that is bound to have limitations, but understanding them can allow the user to use the program in a way that assists with the translation.[31] In order to help researchers and translators select and use CAT systems, the above-mentioned linguistic challenges will be discussed.

Linguistic ambiguity involves situations where words have several meanings, whether based upon multiple definitions of a word (lexical) or the word's placement in a sentence (syntactical).[32] (Discussed further in this section, *infra*). Take for instance, the word *lie*. First, this word is problematic because the word can be classified into two lexical categories: noun (e.g., act of lying) or verb (e.g., to tell an untruth). Such ambiguity would require disambiguation, analyzing the word's placement in the sentence to determine whether the word is being used as a state (i.e., noun) or process (i.e., verb).

Second, both the verb and noun definitions of *lie* are polysemous; here, though, the following example will be limited to the verb form of *lie*. In isolation, the verb form has many different meanings, including "to be or stay at rest in a horizontal position" and to "make an untrue statement with intent to deceive."[33] To decode lexical and syntactical ambiguity, human cognition uses semantic and syntantical cues to determine meaning. Replicating this process in a computer program is difficult, especially for automated translation programs, which historically have had trouble translating texts having semantic ambiguities. Accordingly, a user should identify these ambiguous areas and be prepared to edit a CAT output.[34]

While CAT systems are performing better, common mistranslation problems can still mislead a user. One such issue is a translation that provides an incorrect word that shares the same root as the correct word. For example, *accused* and *accuser* in English share the same root word: *accuse*. An MT output may incorrectly translate the *accuser* to *accused* in the target language. Or, the MT output might not differentiate between the words and provide the root word *accuse* (a verb versus noun). [35] Further, some MT quantitative evaluative measures, like ROUGE and METEOR, permit stemming and would attribute the same meaning to words that share the same root word.[36] Any of these problems might not be caught

---

[30] Kit & Wong, *supra* note 10, at 302.

[31] Kit & Wong, *supra* note 10, at 319-21.

[32] Somers, *supra* note 9, at 612-613.

[33] MERRIAM-WEBSTER ONLINE DICTIONARY (last visited May 21, 2013), http://www.merriam-webster.com/dictionary/lie.

[34] Somers, *supra* note 9, at 612-613.

[35] Email from Harold Somers, Professor of Language Engineering (Emeritus), School of Computer Science, University of Manchester, to Matthew Gran (Dec. 12, 2011).

[36] Jesus Gimenez and Lluis Marquez, *Linguistic Measures for Automatic Evaluation*, 24 MACHINE TRANSLATION 209, 213 (2010). Recall-Oriented Understudy of Gisting Evaluation (ROUGE) is a metrics based program used by MT that relies on n-gram pairings; the evaluation method looks at word pairings as common indications of meaning in translation. *See* Chin Yew-Lin, ROUGE, http://www.berouge.com/Pages/default.aspx (last visited Mar. 9, 2015). Automatic Machine Translation Evaluation System ("METEOR") is a quantative measuring program that scores translation based on different relationships to words and phrases (e.g., exact match, stem, paraphrasing). *See* Michael Denkowski and Alon Lavie, METEOR, http://www.cs.cmu.edu/~alavie/METEOR/ (last visited Mar. 9, 2015).

by an international law researcher/practitioner, and would mislead and possibly confuse the user. For a stemming mistranslation problem, consider the hypothetical criminal case of State v. Stone. When an MT programs translates the State v. Stone opinion, it translates a string into *Stone is the accuser* instead of *Stone is the accused*. This mistranslation could easily be misleading or cause confusion because the mistranslation could suggest Stone accused another person of murder versus being the criminal defendant.

Legal language is also specialized, which affects lexical and syntactical meanings. At the lexical level, legal terms and phrases can have a narrower or broader meaning than in the simple vernacular. Legal terms that are narrow restrict the meaning of the word. Biasiotti and Tiscornia give the example of *worker*. Within employment law, *worker* has a specialized meaning; for example, statutes can tie the definition of *worker* to a list of specific occupations, naturally excluding other jobs. The general meaning of worker, though, is much broader. The opposite can also occur where legal terms have broader meanings rather than general meanings. Here, Biasiotti and Tiscornia give the example of *producer*. Some legal systems include in the definition of *producer* as "those in the supply chain as producers." This definition exceeds the typical general meaning associated with *producer*, which often limits the definition to "the person or entity that made the good." [37] Thus, CAT system users should strive to understand legal terminology and concepts in the language pairing, so they are able to identify whether a translation incorporates the proper lexical and syntactical meanings.

CAT output accuracy is also affected by the degree of difference between languages (i.e., language gap). For instance, MT systems historically had difficulties translating languages that are linguistically distant (e.g., Arabic to English).[38] Recent MT system developments, though, demonstrated that an MT system could translate a linguistically distant language pairing more accurately than a more similar language pairing. For example, the Google Translator performed better in an MT accuracy test from Arabic to English, than from several other languages tested, including several European languages that are more similar to English than Arabic.[39] However, these results are likely more indicative of the amount of U.S. government funds spent on increasing the accuracy of machine translation systems that utilized an English and Arabic language pairing.[40] MT systems, though, theoretically perform better when the languages are linguistically similar.[41]

"Legal language is also specialized, which affects lexical and syntactical meanings."

"CAT output accuracy is also affected by the degree of difference between languages."

---

[37] Maria Angela Biasiotti and Daniela Tiscornia, *Legal Ontologies: The Linguistic Perspective*, in APPROACHES TO LEGAL ONTOLOGIES 143, 162-63 (G. Sartor et al. eds., 2011).

[38] Chunyu Kit and Tak Ming Wong, *Comparative Evaluation of Online Machine Translation Systems with Legal Texts*, 100 LAW LIBR. J. 299, 312 (2008).

[39] Chunyu Kit and Tak Ming Wong, *Comparative Evaluation of Online Machine Translation Systems with Legal Texts*, 100 LAW LIBR. J. 299, 316-17 (2008).

[40] Email from Harold Somers, Professor of Language Engineering (Emeritus), School of Computer Science, University of Manchester, to Matthew Gran (Dec. 12, 2011).

[41] Kit & Wong, *supra* note 39, at 303-306.

Regardless of the language, legal translation is difficult because a legal term or phrase is often rooted in a country's legal traditions.[42] These traditions have caused different legal systems to treat the same legal concept differently. For example, take *defamation*: depending on the legal system, this delict applies to penal law, tort law, or both.[43] Understanding these differences is important to accurately translate legal texts, especially in supranational and international legal systems that require a law to be translated across languages and legal systems. For instance, the European Union may adopt a treaty. A member state may need to adopt the treaty or the treaty may be automatically incorporated into the nation's laws. The European Union law has an intended unified meaning.[44] However, the member states' legal traditions can treat the legal concept addressed in the treaty differently. Here, a translation should be written in a way that expresses the EU meaning.

When working with international and foreign legal texts it is critical to learn and understand these legal traditions, so the translation accurately reflects the law. If a law has three elements in the first language and two in the second language, translators and researchers need to address this difference. Here, an MT output of such a passage would likely require editing and revision, and would certainly not be accurate enough for a government sponsored translation.[45] Further scrutiny must be applied to legal terminology that has different meanings. This terminology misalignment problem can result when the legal term is defined differently in statutes or the term's current definition evolved from multiple legal concepts.[46]

While relying on CAT, legal researchers should consider how the information will be used. Document screening for certain terminology associated with an area of law is different from translating an EU Document from one language to another. Selecting the right CAT for the job often hinges on the user's need.[47] CAT system types will meet the user's needs with various degrees of success. Users who understand the different available CAT programs can select the appropriate translation system that meets their translating needs.

MT systems include those that are rules-based, statistics based, and hybrids of both.[48] Rules-based MT systems translate languages by taking the inputted text and translating the text based upon a series of rules coded into the program. These rules are computer algorithms that are designed to incorporate the lexical, semantic, and syntactical aspects of the language pairs into the MT system.[49] Rules based systems were historically the dominant type of MT, and while they continue to be developed, MT system designers are increasingly developing Statistical MT (SMT) and hybrid based systems.[50]

_____

[42] Alexander Boer et al., *Using Ontologies for Comparing and Harmonizing Legislation*, *in* PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW (ICAIL) 60, 62-63 (2003).
[43] *Id.*
[44] *Id. at 65.*
[45] *Id.* at 62.
[46] GIANMARIA AJANI ET AL., TERMINOLOGICAL AND ONTOLOGICAL ANALYSIS OF EUROPEAN DIRECTIVES: MULTILINGUISM AND THE LAW, 44.
[47] Harold Somers, *The Use of Machine Translation by Law Librarians—A Reply to Yates*, 99 LAW LIBR. J. 299, 302 (2007).
[48] *Id.*
[49] Felipe Sanchez-Martinez and Mikel L. Forcada, *Free/Open-Source Machine Translation: Preface*, 25 MACHINE TRANSLATION 83, 84 (2011).
[50] Mikel L. Forcada at al., *Apertium: Free/Open-Source Machine Translation*, 25 MACHINE TRANSLATION 127, 128 (2011).

Statistics-based MT (SMT) systems are built using a parallel corpus of documents in a language pairing to create a translation model based on the translation work done by human translators. For example, Google Translate is an open source SMT. The statistical program can translate an input based upon prior translations, and selecting outputs based upon higher frequency of similar translations in the corpus. A language model is also created for the source and target languages, by using a computer algorithm that uses statistical probability to analyze the language's linguistic elements.[51] SMT translation accuracy improves with a larger training corpus and also when the terms cover the same subject as the documents to be translated.[52] Therefore, it is more beneficial for a legal researcher/ translator to find SMT's that were built with legal texts. For example, the EU-funded LetsMT! Project built several domains, including a legal corpus of documents between Latvian and English.[53]

However, it is unlikely that most international law researchers and practitioners will use a proprietary SMT exclusively with a legal text corpus because they will likely use open-source SMTs. While open-source SMTs like Google Translate utilize parallel corpora from government sponsored translation by the EU and the UN, these corpora are collected by crawling through the Internet for parallel documents. Google Translate users should recognize the SMT is built upon multiple domains. Accordingly, open-source SMT users should expect these programs to be susceptible to error because the MT output will often include word and phrase translations based upon the most popular definition/translation in the general corpus, instead of accounting for the legal domain's specialized vocabulary.

Lastly there is the hybrid approach that has integrated the statistical and rules based approaches. While SMT improved dramatically over the last decade, MT developers realized that coupling SMT with rules-based MT could help increase translation quality. Hybrid systems use a base line statistical model, but integrate rules based elements into the MT system. A comparable legal analogy to hybrid systems would be that the system uses statistical based algorithms for translation like a general rule of law but applies rules-based MT elements like exceptions, which are designed to account for and correct situations that fall outside the general rule. Hybrid systems can rely on lexical and syntactical translation data encompassed in the RBMT and supplement translation with patterns found by the SBMT part. The development of hybrid systems will hopefully help MT to continue to improve.[54]

----

[51] Sadaf Abdul Rauf and Holger Schwenk, *Parallel Sentence Generation from Comparable Corpora for Improved SMT*, 25 MACHINE TRANSLATION 341, 342-43 (2011).
[52] *Id.* at 343.
[53] *See Lets MT!*, https://www.letsmt.eu/Translate.aspx
(last visited Mar. 9, 2015). Access the *About* section for specific project information.
[54] Andreas Eisle et al, *Hybrid Machine Translation Architectures within and beyond the EuroMatrix project*, 12TH EAMT CONFERENCE, 27, 27-30 (Sep. 2008).

Given the above analysis, you may be wondering about how you can effectively apply this information to the use of MT for legal texts. Scholars and practitioners evaluated plenty of MT systems over the years: using qualitative (user-focused) and quantitative methodology (applying a predetermined measurement). For instance, Somers proposed applying qualitative measures that rank translations based upon their usefulness to perform a task. For example, a law librarian using an MT system to translate a cataloguing record is a very different task from a professional translator trying to translate an entire international treaty.[55] An MT output may be beneficial and help the law librarian catalog the foreign book; however, the machine translated treaty would likely not assist the professional translator.[56]

Users can also spot check an MT, especially when they know of terms and phrases that suffer from lexical and syntactical ambiguity. Does the system accurately translate a legal term with a broader meaning than the vernacular? Can the system translate legal terms that may be collocated in one language, but separated in another? While a qualitative measurement approach has not been formally applied in studies, a forum or wiki could be an excellent place to share reviews of CAT. Shared websites for legal translators could include evaluations of CAT applications, and could specifically address how well they handle linguistic issues.

Historically, the standard MT quantitative evaluation is the BLEU score, developed by IBM, which focuses on the degree of precision between the source and target language (i.e., how closely the MT output overlaps with a human created model translation). [57] The "BLEU" score calculates a sentence score based on the percentage of the output matches between MT and a human translation. A good BLEU score will approach 1.0, while a lower score will approach 0.0.[58] The test emphasizes the correct choice of words, but does not factor in the order of words, and it penalizes translations that have fewer words in the translation than the original document. This measure has been criticized because longer translations may capture the semantic and lexical meaning better than a shorter sentence.[59] Regardless, a user can look to qualitative evaluations of MT outputs as approximations of how well an MT system performs, but should remember that a system can perform better depending upon the domain of document(s) to be translated. Additionally, an MT system may translate one language pairing well and not another.[60]

_____

[55] Harold Somers, *The Use of Machine Translation by Law Librarians—A Reply to Yates*, 99 LAW LIBR. J. 611, 618, 619 (2007).
[56] *See id*.
[57] Email from Harold Somers, Professor of Language Engineering (Emeritus), School of Computer Science, University of Manchester, to Matthew Gran (Dec. 12, 2011).
[58] Kishore Papineni et al., *BLEU: a Method for Automatic Evaluation of Machine Translation*, PROCEEDINGS OF THE 40TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL) 311, 312-315 (July 2002).
[59] *See* Chunyu Kit and Tak Ming Wong, *Comparative Evaluation of Online Machine Translation Systems with Legal Texts*, 100 LAW LIBR. J. 299, 309.
[60] *Id.*

Overall, translation programs, including MT systems, are becoming better tools to assist international law practitioners, researchers, and translators. The user's intended use still largely factors into the usefulness of MT and other translation systems' applications to legal texts. An MT system may be great for those seeking a "quick and dirty translation," but users may opt for an MT corpus that is limited to legal texts for a more precise translation. Increasingly, scholars are suggesting that professional translators should utilize computer-assisted translation systems to increase translation efficiency and performance. These translations use parallel corpora and demonstrate several sentences, each with the word or phrase to be translated. Another option is using translation memory systems, getting the benefit of reusing prior professional translations, but being limited to words and phrase translation. There are also free concordances, like [Linguee.com](http://www.linguee.com) (http://www.linguee.com), that help provide various examples of translated text. Users who consider the various linguistic challenges involved in translation, MT system challenges, their translation needs, source and target language proficiencies, and various translation systems are better equipped to select an appropriate translation system. Ultimately, systems cannot create a great translation on their own, but they will help increase translators' effectiveness.

Workplace MT Usage: Examples

Drawing now on Part I and II of this article, we would like to show a few simple examples of machine translation (MT). This is not a statistical study or survey, but simply some illustrations of how information professionals may use MT with a modicum of confidence.

As discussed previously, MT will work best in a legal context when languages are used that have built up relatively large dictionaries. This will occur with larger languages (Chinese; Indonesian; Russian), and particularly with large "interlanguages" (English and Spanish) or major traditional interlanguages (such as French, particularly in a diplomatic or international law context). Above all, a realistic use of MT would be done by someone going from a language in which they have good command into the language they consider their first language (or "mother tongue"). Finally, having a command of the specific field in question, and its terms of art (or knowing how to successfully research its terms of art) is critical. Thus, in our case, the user of MT should have training in the law and in legal research.

Translations of bibliographic records are increasingly doable with MT. Please be aware, however, that even with bibliographic records, the MT must always be supplemented by good print or electronic legal/business dictionaries, both in the vernacular exclusively, and in English/vernacular editions. Whenever possible, try to translate phrases, sentences, or possibly fields (like a title field in a bibliographic record). The translation of entire paragraphs is sometimes doable, increasingly so when the translation tool is custom-designed for the database (like the Spanish/English translations on vLex) or are harvesting governmental multilingual document archives (as done by the United Nations, the EU, and companies like Linguee, which harvest UN and EU documents and build dictionaries accordingly).

Following are four examples of bibliographic record translations using German, Spanish, Indonesian (Bahasa Indonesia), and Chinese (Pinyin) examples.

First, a German legal title and its variations according to several MT tools. The title is:

Der O.-J.-Simpson-Prozess: das strafprozessuale Vorverfahren und dessen Auswirkungen in rechtsvergleichender Darstellung

Here is a translation using both MT and print sources:

The O.J. Simpson Trial: the criminal pre-trial proceedings and their implications in a comparative law presentation.

Here are several results using MT exclusively:

Systran:

Google Translate:

Google translation tool used to directly translate a page from the item's OPAC record at the German Na-

tional Library (Deutsche Nationalbibliothek):

German record:

English translation using the Google translate tab:

All of the MT tools yield relatively good results. The actual translation of the bibliographic record, not being done with the translation tool window, was still accurate in comparison. This was probably due to the fact that the item is organized in fields.

Another example, from the world interlanguage Spanish, for which huge and exponentially growing MT dictionaries exist due to the demand for translating between the Spanish- and English-speaking worlds. Here is a straightforward title dealing with MERCOSUR[61] and a unique Latin American remedy (something along the lines of writs in the common law tradition[62]):

> **MERCOSUR: la protección de los datos personales: privacidad versus derecho a la información, regimen legal en el MERCOSUR y en la Unión Europea, habeas data**

Again, a translation that combines the use of MT and print resources:

> **MERCOSUR: protection of personal data: privacy versus the right to information, the legal framework of MERCOSUR and the European Union, [remedy or procedure of] habeas data.**

Here are several MT renditions of the title:

Systran:

---

[61] Mercado Común del Sur in Spanish; Mercado Comum do Sul in Portuguese.

[62] The remedy of "habeas data," found in countries such as Brazil, Paraguay, Argentina, and the Philippines.

Google Translate:

Now, using the Google tab for translation for the relevant webpage itself:

Spanish language version of the title in the bibliographic record:

Now, using the translation tab:

MERCOSUR : la protección de los datos personales : privacidad versus...

Relevance:	⬤ ⬤
LC control no.:	00279751
LCCN permalink:	http://lccn.loc.gov/00279751
Type of material:	Book (Print, Microform, Electronic, etc.)
Personal name:	Slavin, Diana de.
Main title:	MERCOSUR: the protection of personal data: privacy versus right to information, legal regime in MERCOSUR and the European Union, habeas data / Diana Slavin, prologue, George Beckerman.
Published / Created:	Buenos Aires: Ediciones Depalma, 1999.
Description:	xiii, 167 p. , 20 cm.
ISBN:	9501417360
CALL NUMBER:	KH103.C65 S59 1999

The Spanish language translation is remarkably accurate. The basic free MT tools seem to work well for European languages such as German and Spanish, and are a great help if the user already has some facility in the languages, and a training in the subject matter of the titles.

However, what would the results be with languages that move away from a common European/Latin origin? Below are two examples from Southeast Asia, one using Indonesia's national language of Bahasa Indonesia. The second involves the Pinyin version of Chinese.

Bahasa Indonesia ("Language of Indonesia") is based on the Malay language and in 1928 was declared the national language of an emerging Indonesia independent of Dutch colonial rule. In the 1945, the first post-independence constitution recognized Bahasa Indonesia as the official language. The characters are based on the Roman alphabet as used by the Dutch colonial rulers, although an alphabet reform took into account English usage because of the latter's use as a major world interlanguage.

Here is a bibliographic title in Bahasa Indonesia:

ar
s by: | *Year* | Author | Title | Call Number | Format

◆ [Display full record] [Display Availability]

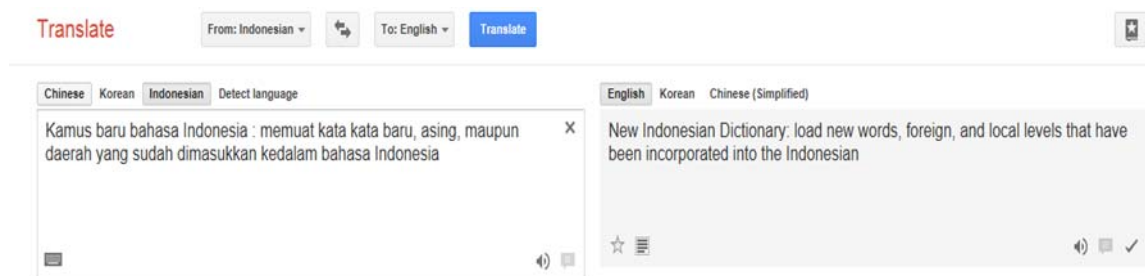|	
Title	Kamus baru bahasa Indonesia : memuat kata kata baru, asing, maupun daerah yang sudah dimasukkan kedalam bahasa Indonesia /
Published	Surabaya : Usaha Nasional, 1984.
Format	Book
Location	Main Library PL5077 .K32 1984
Availability	Check shelf or Request delivery

Using Google Translate we get:

This title translation is fairly accurate, and could be translated easily by someone with facility in Bahasa Indonesia and working with a dictionary, either electronic or print. By looking merely at the Google Translate result, one might think that "load" is a mistaken translation for "loan" and conclude that loan words are a particular focus of this dictionary. However, while "memuat" can refer to "load," it can also mean "contain" or publish [the contents of something]. However, "kata kata baru" does refer to new words. However, Google Translate lists "loan words" as "kata pinjaman" and "pinjaman" itself "loan" and "pinjam" as "borrow." Publication of new words might be include loan words. For bibliographic record purposes a user not familiar with Bahasa Indonesia might nonetheless attempt a translation with something as relatively simple as a bibliographic title for the simple reason that Bahasa Indonesia takes much of its structure from Dutch and English.

However, what would happen if Chinese Pinyin were used instead? Pinyin is basically a Romanized Chinese. However, the underlying structure of Chinese itself has not been affected by European languages as intensively as Bahasa Indonesia has been. Here is a
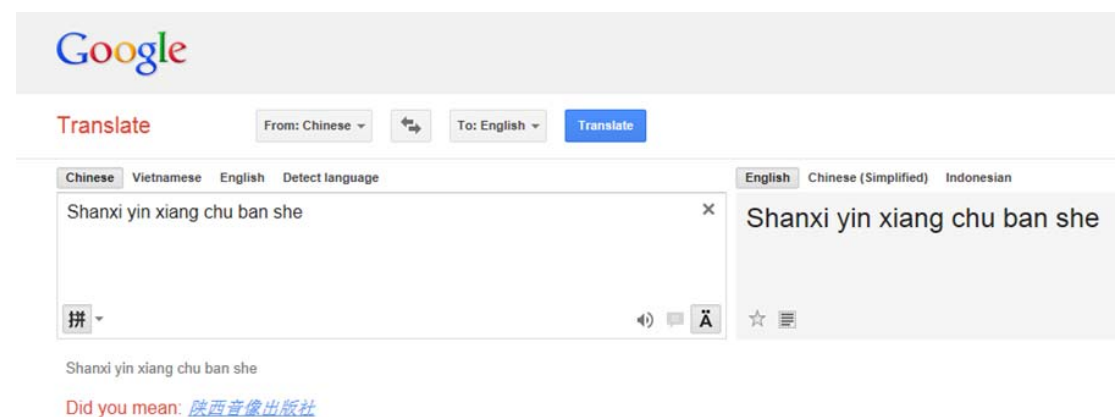

basic Pinyin title:

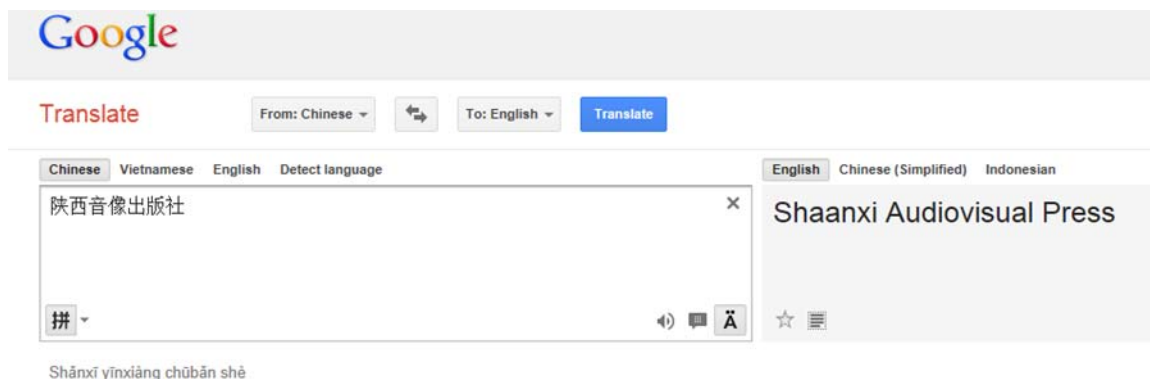Using Google Translate, we get the following result for the title:
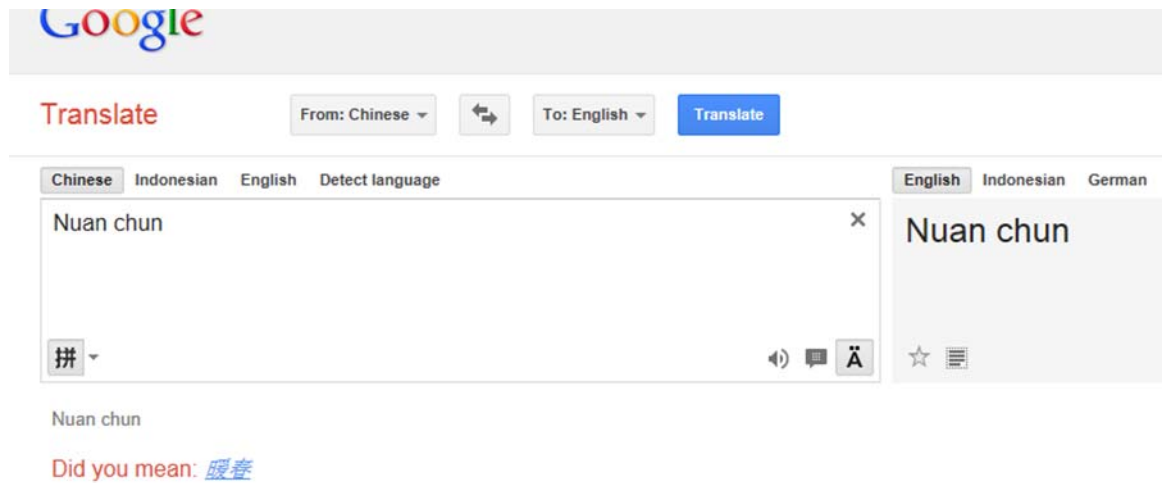
By clicking on the Chinese characters in the lower left hand corner where it says "Did you mean[,]" we get:
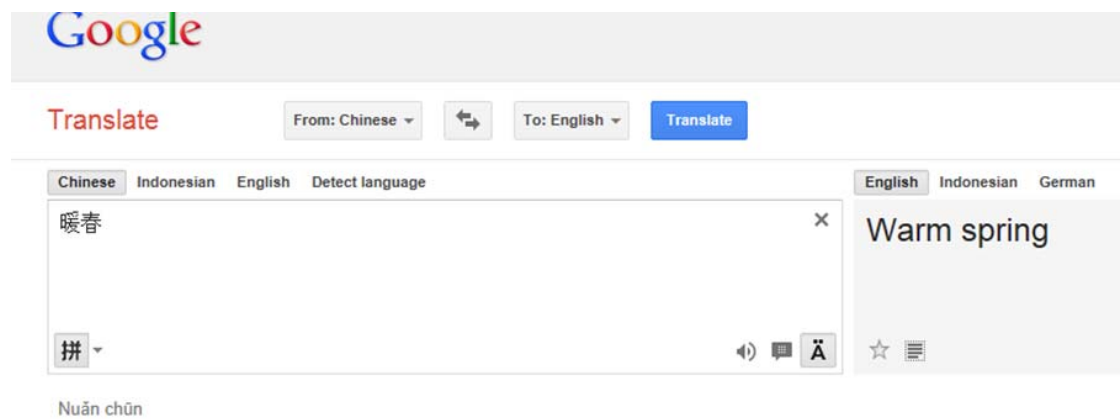
This is the same title found on the Internet Movie Database (IMDb).

Again, clicking on the "Did you mean" link in the lower left hand corner, we get:

Likewise, again the results are not poor, but "Shanxi" the province comes out as "Shaanxi" in the translation. However, since the translation involves a DVD, the term "Audiovisual Press" seems in keeping with the medium.

However, author Ford tried translating several other phrases from the "Published" field with varying degrees of success. In the case of Pinyin, one would either want to have fluency or capability in Chinese, or would want to be searching for the absolute minimum gist of something.

Conclusion

MT and the web are now making it possible to craft accurate translations, particularly between strong interlanguages (such as Spanish and English), which have built up robust MT dictionaries. But the "crafting" of the translation still needs an adept user of both the original and target languages. And remember: a truly good translation involving the combination of MT and language dictionaries (either electronic or print), only reaches its high level of accuracy when the translator is not only knowledgeable with the languages, but also with the specialized subject matter, such as law.

Selected Bibliography of Machine Translation Resources

Monographs

Alcaraz Varó, Enrique. *Legal Translation Explained*. Manchester, U.K.: St. Jerome Publishing, 2002.

Cao, Deborah. *Translating Law*. Buffalo, New York: Multilingual Matters, 2007.

Carl, Michael, and Andy Way (editors). *Recent Advances in Example-Based Machine Translation*. Boston: Kluwer Academic Publishers , 2003.

Chan, Sin-wai. *A Topical Bibliography of Translation and Interpretation: Chinese-English/English-Chinese*. Hong Kong: Chinese University Press, 1995.

Gerzymisch-Arbogast, Heidrun, and Helle V. Dam and Jan Engberg (editors). *Knowledge Systems and Translation*. New York: Mouton de Gruyter, 2005.

Nirenburg, Sergei, and Harold L. Somers and Yorick A. Wilks (editors). *Readings in Machine Translation*. Cambridge, M.A.: MIT Press, 2003.

Translation Research and Interpreting Research: Traditions, Gaps and Synergies. Buffalo, NY: Multilingual Matters, 2004.

Translations of Commonly Used Court Related Phrases. Chicago: American Bar Association, 1998.

Wilks, Yorick. *Machine Translation: Its Scope and Limits*. New York: Springer, 2009.

Wilss, Wolfram. *Knowledge and Skills in Translator Behavior*. Philadelphia: John Benjamins Publishing, 1996.

Periodicals

Kit, Chunyu, and Tak Ming Wong. "Comparative Evaluation of Online Machine Translation Systems with Legal Texts." *Law Library Journal* 100 (Spring 2008): 299-321.

Sexton, Mary. "Preferably in English: Surfing the Pacific in Search of Law in Translation." *Syracuse Journal of International Law & Commerce* 35 (Spring 2008): 275-291.

Somers, Harold. "The Use of Machine Translation by Law Librarian—A Reply to Yates." *Law Library Journal* 99 (Summer 2007): 611-619.

Yates, Sarah. "Scaling the Tower of Babel Fish: An Analysis of the Machine Translation of Legal Information." *Law Library Journal* 98 (Summer 2006): 481-500.

International Legal Research Group

The International Legal Research Interest Group (ILRIG) is dedicated primarily to its members' professional development in the areas of foreign, comparative, and international law (FCIL). ILRIG provides a forum for discussion among legal information professionals, legal scholars, and attorneys. ILRIG enhances its members' opportunities to share their knowledge about available FCIL resources, research methods, research techniques, and best practices. ILRIG organizes presentations, publishes a newsletter, and maintains a website that reflects the most recent developments in the legal research profession.

ILRIG members are particularly mindful of the interdisciplinary and multicultural aspects of contemporary foreign, comparative, and international law. Global legal policies and norms cannot exist without strong foundations built on exhaustive research. ILRIG is committed to being a forum for discussing ASIL's unique analytical needs.

ILRIG membership is open to all ASIL members. ILRIG should be of particular interest to:

- Law librarians
- Legal scholars
- Attorneys with FCIL practice issues
- Academic librarians
- Scholars working in political science, international relations, economics, and history
- Research professionals from government agencies, policy institutes, inter-governmental organizations, and non-governmental organizations

ILRIG OFFICERS

Co-Chairs
Jootaek "Juice" Lee
Senior Law Librarian & Affiliated Faculty of Program of Human Rights and Global Economy, Northeastern University School of Law

Wanita Scroggs
International Law Librarian and Adjunct Professor, Stetson University College of Law

Secretary
Marylin Raisch
Associate Law Librarian for International and Foreign Law, Georgetown Law School

Treasurer
Gabriela Femenia
Foreign and International Law Librarian, University of Pennsylvania Law School

Newsletter Editor(s)
Don Ford
Foreign, Comparative, and International Law Librarian, University of Iowa College of Law

The Informer

The *Informer* is the bi-annual newsletter of the International Legal Research Interest Group (ILRIG). Any views expressed in this newsletter are those of the authors in their private capacities and do not purport to represent the official view of the ASIL or ILRIG.

Submissions are welcomed and will be published at the discretion of the editors. Essays or articles should relate to foreign, comparative, and international law (FCIL) resources, research methods, research techniques, and best practices.

To contribute to future issues of the *Informer*, contact:

- Don Ford at donald-ford@uiowa.edu
- Jootaek ("Juice") Lee at joo.lee@neu.edu