

# Computational Linguistics

A First Academic Investigation -  
From General to More Specific and Academic Content

1/30/2010

Bamshad Lotfabadi

An Investigation for the Research Methodology Course,  
Allameh Tabataba'i University, Tehran, Iran  
Professor: Dr. G. R. Tajvidi

# Computational Linguistics

*A First Academic Investigation -  
From General to More Specific and Academic Content*

This investigation is consisted of five inter-related sections which make the overall body of my research on Computational Linguistics: 1. General Introductory Articles, 2. Master's Programs Offered by Credible Universities in the Field, 3. Definitions and General Terms – Computational Linguistics, 4. Detailed and Field Specific Articles and Content, and 5. Further Graduate Studies in Detail.

## Content

### General Introductory Articles

1. [Page 4] - The National Science Foundation
2. [Page 4] - The Distinguished Alumni Award
3. [Page 5] - The Association for Computational Linguistics
4. [Page 5] - The Language Software Revolution

### Master's Programs Offered by Credible Universities in the Field

5. [Page 8]
6. [Page 9] - Master's Program in Computational Linguistics – University of Washington
7. [Page 9] - Information Sciences Institute - Master's Program in Computational Linguistics - University of Southern California
8. [Page 10] - Computational Linguistics – Department of Computer Science - University of Toronto

### Definitions and General Terms – Computational Linguistics

9. [Page 11] - Definition of Computational Linguistics – Wikipedia.com
10. [Page 11] - Artificial Intelligence – Wikipedia.com
11. [Page 13] - Intelligence – Wikipedia.com
12. [Page 15] - Intelligence – Wikipedia.com
13. [Page 16] - Language Technology
14. [Page 19] - Machine translation (MT), and the future of the translation industry

### Detailed and Field Specific Articles and Content

15. [Page 21] - XML and the Translator
16. [Page 24] - Open source translation management system
17. [Page 26] - XML-Intl.
18. [Page 27] - Automatic Translation in Multilingual Electronic Meetings
19. [Page 30] - Statistical natural language processing and corpus-based computational linguistics: An annotated list of resources
20. [Page 30] - Book Review: An Introduction to Language Processing with Perl and Prolog
21. [Page 31] - Book Review: Computational Linguistics: Hammond (2003)
22. [Page 32] - Computational Resources for Linguistic Research

### Further Graduate Studies in Detail

23. [Page 33] - Information Sciences Institute - Master's Program in Computational Linguistics - University of Southern California – Courses Offered
24. [Page 36] - University of Southern California – Master's Program in Computational Linguistics
25. [Page 37] - Graduate Studies in Computational Linguistics – University of Toronto – Courses Offered



## Part One: General Introductory Articles

### 1.

[http://www.nsf.gov/news/news\\_summ.jsp?cntn\\_id=114063](http://www.nsf.gov/news/news_summ.jsp?cntn_id=114063)

## **The National Science Foundation**

The National Science Foundation (United States) has planned an Olympiad in Computational Linguistics for high school students across North America. The National Science Foundation (NSF), being an independent governmental organizations provides aid for simply any of the branches of science across the United States with 3 billion dollars of funding resource.

Students will be given puzzles and translation-related problems to deal with. According to the Article, Computational Linguistics has significant importance because of the emerging global threats and technological advancements by other countries. The aim of the Olympiad is to offer a chance to all high school students to test their talents in an ever-increasing popular field of science in the world

### 2.

<http://newsinfo.iu.edu/news/page/normal/10567.html>

## **The Distinguished Alumni Award**

Lauri Karttunen will receive the Distinguished Alumni Award from the IU Department of Linguistics.

Karttunen said he has spent his entire career trying to build systems that understand natural spoken language, the ultimate goal of computational linguistics.

"One indication that a human has understood a piece of text is that she can answer questions about it," Karttunen said. "A computer should be able to do the same. It is unlikely that we will get there in my lifetime, but progress is being made."

### 3.

[http://en.wikipedia.org/wiki/Association\\_for\\_Computational\\_Linguistics](http://en.wikipedia.org/wiki/Association_for_Computational_Linguistics)

## The Association for Computational Linguistics

The Association for Computational Linguistics is the international and professional society for people working on problems related to Natural Language and Computation

### 4.

<http://www.translationdirectory.com/articles/article2037.php>

## The Language Software Revolution

By Dr. Derek Mohammed, PhD., Assistant Professor,  
Our Lady of the Lake University,  
San Antonio, Texas USA

The continued expansion of the global market, and the realization that English may have already reached its zenith as the global internet language, requires transnational corporations to utilize multilingual means of reaching new markets. In the process of accessing new audiences for their products, transnational corporations must ensure that the messages in marketing ads are grammatically and culturally correct. In addition, new translation programs are currently being developed that will provide transnational organizations the ability to instantly communicate in the multilingual environment of the global marketplace, blurring even more the political and cultural borders of our world. Furthermore, these new programs will affect every aspect of our social and cultural environments, and will add both to the fragmentation and the globalization process underway.

Globalization creates an illusion that the Earth is shrinking due to the power of technology to interconnect world markets. Technology provides the means for companies to do business rapidly with the click of a mouse and the change from domestic to global markets will create a paradigm shift in how business competes in the global economy. The ability to communicate via the internet with almost any nation on the planet increases the capacity for business to participate in global markets and with this expansion of business has deep ramifications that create the opportunity for complete shifts in the mindset and practices of business.

...

Because machine translation alone does not provide the type of accuracy that is needed due to its failure to recognize the subtleties within languages, various companies are attempting to address this deficiency. A system that has shown promise is Synchronous Automated Translation System (SATS). The system is intended to provide instantaneous translation while also providing proper syntax, idiom phonetics and spelling (Lehman-Wilzig, 2001). The program will be networked

thus providing a seamless translation between speakers. Additionally, a dictionary module tied into the system will provide for instant definitions, and the information will automatically update itself as long as the user is online. The ultimate objective of an online SATS program will be the ability to translate phone, video, radio, and Internet information into the language of the user. Additionally, whereas professional translators can effectively translate three to five languages, the SATS unlimited memory and online networking will enable it to translate many more languages.

Similar to SATS is xTALK (crosstalk), translation software created at the National University of Singapore. This system is a combination of voice recognition, machine translation and speech synthesis technologies which provides for speech translation. This system currently has a six second delay between the time a person speaks to the time the system performs the translation. The system continuously improves as all translations are cached thus new translations can be provided from the cached material. Voice recognition capabilities of the system include English, Chinese and Japanese, while speech synthesis capabilities of the system include English, Chinese, Japanese, French, Spanish and German. By incorporating speech synthesizing capabilities of existing providers, the system has the potential to provide international conferencing and instant messaging.

Very similar in function to xTALK is Phraselator, a handheld speech translator used by the U.S. military and provided by Tec International. The Phraselator incorporated the capabilities of a PDA thus allowing for the translation capabilities. By speaking into the Phraselator, the device then translates the phrase or sentence and will provide the translation through a speaker. It is limited to approximately 1,000 phrases that can be translated into 40 languages. The U.S. military found a great need for such a device by virtue of the many nations in which it operates. It saw the need to engage with non-English speakers in order to more effectively interact and maintain order. Similar in nature, and already in use is the e-NAVI system utilized by the Japanese. The system is in effect a PDA that provides English to Japanese translation for tourists. The e-NAVI system is set up to allow for translation, phone service, planner, guide, and illustrated brochure, and access to the Internet. As previously noted, more advanced capabilities will include the future use of translation devices that will be able to translate and process language without human interaction. A cell phone user can converse in their language as they speak to a person speaking in a different language through the use of a computer that translates the languages in both directions. Current capabilities of this system are limited to English, German and Japanese with an accuracy of between 80 to 90 percent. According to Belluomini (2006) the ultimate intent of multilingual communications is being able to not only translate, but provide cultural interpretation as well.

Beyond military use of the Phraselator, and the tourist assistance capabilities of the e-NAVI, such devices may see practical use in a number of other areas, law enforcement agencies in large American urban centers, as well as urban centers throughout the world where large concentrations of non-native speakers reside, can make use of translators/PDAs. In the international arena where political differences can lead to conflict, the use of machine translators can provide instant communication that may assist in diffusing possible hostilities. Such devices can also assist in the health care field where communication between doctor and patient is imperative. As with tourists, machine translators can also assist newly arrived immigrants whose limited knowledge of the language (and culture) are obstacles to employment, travel, shopping,

and as noted, health care. The entertainment trade, particularly the movie industry may be greatly affected as use of the machine translation will provide proper and instant translation of the movie's dialogue.

As with any new items on the market, the system being developed come with high price tags. Phraselators range in price from approximately \$3,200 to \$1,200. Lingo Voyager 2 by Lingo Corp., produces a PDA translator similar to the Phraselator that costs approximately \$200. Lingovobit Inc. also sells a translator known as the SpeechGuard TL 2m5 for approximately \$450. The cost for larger integrated systems such as xTALK and SATS depend on the extent to which the systems are incorporated into a company's electronic infrastructure.

Through the application of the systems noted above, corporations will be able to engage in instance teleconferencing and exchange information. Electronic commerce will be greatly enhanced by virtue of instant translation and clarification of desired transactions. In effect the language barriers to further globalization are being surmounted by language technology. However, machines will never replace the ambiance provided by human interactions. From the perspective of the authors, global leadership still requires human leadership, and no quality of translation can ever replace this.

Perhaps it is important to keep in mind two perspectives provided by authors whose works were consulted. With regards to the globalization process and continued improvement of machine translations: "Automated translation systems may be either a centrifugal factor that fragments the world, or a centripetal force that binds cultures closer together. Both trends could well occur simultaneously" (Lehman-Wilzig, 2001). And with regards to a single world language: "Many feel that while a single international language would make personal and business interactions easier, more would be lost than gained. Monocultures can kill ecologies of ideas – and new ideas are the most important asset of any business" (Grossman, 2000).



## Part Two: Master's Programs Offered by Credible Universities in the Field

## 5.

## 6.

<http://www.compling.washington.edu/compling/>

### **Master's Program in Computational Linguistics – University of Washington**

The UW Department of Linguistics, one of the oldest in the U.S., brings you a top-notch Professional Master's program in Computational Linguistics. Steered by an advisory board that includes managers and researchers from many leading technology companies, our program is designed to ensure you'll have the skills to grow and change with this exciting, dynamic field.

#### **What is Computational Linguistics?**

Language is widely recognized as part of what makes us human. We instinctively know its value in communication and the development of ideas. Transferring those skills to computers is the challenge of a computational linguist—from predictive text messaging to dialogue software for your car, to discoveries in medical research.

In just 12 months of full-time study, or 24-36 months of part-time study, you'll be well prepared for a variety of positions:

- Computational Linguist
- Specialized Software Development Engineer
- Language Technician
- Language Specialist
- Quality Assurance Analyst
- Translational Technology Specialist

## 7.

<http://www.isi.edu/natural-language/MSCompLing/>

### **Information Sciences Institute - Master's Program in Computational Linguistics - University of Southern California**

USC offers two programs in the area of Computational Linguistics (also called Natural Language Processing and Human Language Technology):

- This program offers an MS degree. It is centered in the Department of Linguistics and focuses on issues in Linguistics. Its faculty are primarily experts in Linguistics.
- The other program ([click here](#)) offers MS and PhD degrees in Computer Science with an emphasis on Human Language Technology / Computational Linguistics, and focuses on all aspects of computational linguistics. Its faculty are primarily experts in Computer Science, and are members of the world-renowned Natural Language research groups at the Information Sciences Institute (ISI) and Institute for Creative Technology (ICT).

### **Research Areas**

The faculty members comprise a group of internationally renowned scholars from [Linguistics](#). Their areas of research help to shape research in the program. Within the field of Linguistics, prevailing models of formal grammar have been shaped to a great extent by these scholars whose expertise encompasses not only grammatical theory but a wide range of languages and language families. [Jean-Roger Vergnaud](#) established the notion of Abstract Case theory, which has remained one of the central concerns of formal grammatical theory for more than two decades. [Joseph Aoun](#) pioneered the study of Generalized Binding and has remained in the forefront of investigations of application of formal grammatical models to East Asian and Semitic languages. [Bonnie Glover Stalls](#) is working on automated information extraction for web-based resume text processing. She has developed Arabic lexical, syntactic, and semantic resources for broad-coverage Arabic-to-English machine translation system at ISI and has extensive industrial experience in multilingual applications including voice-to-voice translation. n Computer Science [Michael Arbib](#) is active in computational and cognitive neuroscience as well as neuroinformatics. A topic of current interest relates the mechanisms of control of hand movements in monkeys to a new scenario for the origin of human language. [Shrikanth Narayanan](#) is a member of the Electrical Engineering Department after working at AT&T Research. He focuses on automated speech recognition and speech synthesis.

## **8.**

<http://www.cs.utoronto.ca/compling/index.html>

## **Computational Linguistics – Department of Computer Science - University of Toronto**

### Overview

Computational linguistics is the study of computer processing, understanding, and generation of human languages. It is often regarded as a subfield of artificial intelligence. Techniques from computational linguistics are used in applications such as machine translation, speech recognition, information retrieval, intelligent Web searching, and intelligent spelling checking.

The Computational Linguistics research group of the [Department of Computer Science, University of Toronto](#), carries out research on many topics in the area.

The group's research is supported by research grants and scholarships from the [Natural Sciences](#)

[and Engineering Research Council of Canada](#), from [Communications and Information Technology Ontario](#), from [Bell University Laboratories](#) and by Ontario Graduate Scholarships.

[http://en.wikipedia.org/wiki/Computational\\_linguistics](http://en.wikipedia.org/wiki/Computational_linguistics)



### Part Three: Definitions and General Terms – Computational Linguistics

## 9.

[http://en.wikipedia.org/wiki/Computational\\_linguistics](http://en.wikipedia.org/wiki/Computational_linguistics)

### **Definition of Computational Linguistics – Wikipedia.com**

Computational Linguistics is an interdisciplinary field of science related to statistical or rule-based understanding of natural language field. CL (Computational Linguistics) was considered earlier as a section of Artificial Intelligence, however the recent changes in technological and cultural trends have opened paths for experts of many fields to contribute to the expansion of CL knowledge. As time passed scientists realized that CL explorations are not as simple as they thought it to be since language is a fundamental part of the human nature and is indeed very difficult to analyze.

Computational Linguistic researches can be divided into several major categories:

**One**, Computational Complexity, of Natural Language, largely modeled on Automata Theory, with the application of Context-sensitive grammar and linearly-bounded Turing machines. **Two**, Computational semantics, comprised of defining suitable logics for linguistics meaning representation, automatically constructing them and reasoning with them. **Three**, Computer-aided Corpus Linguistics; **Four**, Design of parsers or chunkers for natural languages. **Five**, design of taggers like POS-taggers (Part of Speech Taggers), **Six**, Machine translation, as one of the earliest and least successful applications of Computational Linguistics, which draws on many subfields.

## 10.

[http://en.wikipedia.org/wiki/Artificial\\_intelligence#Philosophy](http://en.wikipedia.org/wiki/Artificial_intelligence#Philosophy)

## Artificial Intelligence – Wikipedia.com

**Artificial intelligence (AI)** is the [intelligence](#) of machines and the branch of [computer science](#) which aims to create it. Textbooks define the field as "the study and design of [intelligent agents](#),"[\[1\]](#) where an intelligent agent is a system that perceives its environment and takes actions which maximize its chances of success.[\[2\]](#) [John McCarthy](#), who coined the term in 1956,[\[3\]](#) defines it as "the science and engineering of making intelligent machines."

AI research is highly technical and specialized, deeply divided into subfields that often fail to communicate with each other.[\[10\]](#) Subfields have grown up around particular institutions, the work of individual researchers, the solution of specific problems, longstanding differences of opinion about how AI should be done and the application of widely differing tools. The central problems of AI include such traits as reasoning, knowledge, planning, learning, communication, perception and the ability to move and manipulate objects.[\[11\]](#) General intelligence (or "[strong AI](#)") is still a long-term goal of (some) research.[\[12\]](#)

### Problems:

#### (1) [Deduction, reasoning, problem solving](#)

Early AI researchers developed algorithms that imitated the step-by-step reasoning that humans use when they solve puzzles, play board games or make logical deductions.[\[39\]](#) By the late 1980s and '90s, AI research had also developed highly successful methods for dealing with [uncertain](#) or incomplete information, employing concepts from [probability](#) and [economics](#).[\[40\]](#)

For difficult problems, most of these algorithms can require enormous computational resources — most experience a "[combinatorial explosion](#)": the amount of memory or computer time required becomes astronomical when the problem goes beyond a certain size. The search for more efficient problem solving algorithms is a high priority for AI research.[\[41\]](#)

Human beings solve most of their problems using fast, intuitive judgments rather than the conscious, step-by-step deduction that early AI research was able to model.[\[42\]](#) AI has made some progress at imitating this kind of "sub-symbolic" problem solving: [embodied agent](#) approaches emphasize the importance of [sensorimotor](#) skills to higher reasoning; [neural net](#) research attempts to simulate the structures inside human and animal brains that gives rise to this skill

#### (2) [Knowledge representation](#)

Main articles: [Knowledge representation](#) and [Commonsense knowledge](#)

[Knowledge representation](#)[\[43\]](#) and [knowledge engineering](#)[\[44\]](#) are central to AI research. Many of the problems machines are expected to solve will require extensive knowledge about the world. Among the things that AI needs to represent are: objects, properties, categories and relations between objects;[\[45\]](#) situations, events, states and time;[\[46\]](#) causes and effects;[\[47\]](#) knowledge about knowledge (what we know about what other people know);[\[48\]](#) and many other, less well researched domains. A complete representation of "what exists" is an [ontology](#)[\[49\]](#) (borrowing a word from traditional [philosophy](#)), of which the most general are called [upper ontologies](#).

**Among the most difficult problems in knowledge representation are:**Default reasoning and the qualification problem

Many of the things people know take the form of "working assumptions." For example, if a bird comes up in conversation, people typically picture an animal that is fist sized, sings, and flies. None of these things are true about all birds. [John McCarthy](#) identified this problem in 1969[50] as the qualification problem: for any commonsense rule that AI researchers care to represent, there tend to be a huge number of exceptions. Almost nothing is simply true or false in the way that abstract logic requires. AI research has explored a number of solutions to this problem.[51]

The breadth of commonsense knowledge

The number of atomic facts that the average person knows is astronomical. Research projects that attempt to build a complete knowledge base of commonsense knowledge (e.g., [Cyc](#)) require enormous amounts of laborious ontological engineering — they must be built, by hand, one complicated concept at a time.[52] A major goal is to have the computer understand enough concepts to be able to learn by reading from sources like the internet, and thus be able to add to its own ontology.

The subsymbolic form of some commonsense knowledge

Much of what people know is not represented as "facts" or "statements" that they could actually say out loud. For example, a chess master will avoid a particular chess position because it "feels too exposed"[53] or an art critic can take one look at a statue and instantly realize that it is a fake.[54] These are intuitions or tendencies that are represented in the brain non-consciously and sub-symbolically.[55] Knowledge like this informs, supports and provides a context for symbolic, conscious knowledge. As with the related problem of sub-symbolic reasoning, it is hoped that situated AI or computational intelligence will provide ways to represent this kind of knowledge.[55]

**(3) Planning**

Main article: [Automated planning and scheduling](#)

Intelligent agents must be able to set goals and achieve them.[56] They need a way to visualize the future (they must have a representation of the state of the world and be able to make predictions about how their actions will change it) and be able to make choices that maximize the utility (or "value") of the available choices.[57]

In classical planning problems, the agent can assume that it is the only thing acting on the world and it can be certain what the consequences of its actions may be.[58] However, if this is not true, it must periodically check if the world matches its predictions and it must change its plan as this becomes necessary, requiring the agent to reason under uncertainty.[59]

Multi-agent planning uses the cooperation and competition of many agents to achieve a given goal. Emergent behavior such as this is used by evolutionary algorithms and swarm intelligence. [60]

# 11.

<http://en.wikipedia.org/wiki/Intelligence>

## Intelligence – Wikipedia.com

A second definition of intelligence comes from "[Mainstream Science on Intelligence](#)", which was signed by 52 intelligence researchers in 1994:

A very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience. It is not merely book learning, a narrow academic skill, or test-taking smarts. Rather, it reflects a broader and deeper capability for comprehending our surroundings—"catching on", "making sense" of things, or "figuring out" what to do.

Researchers in the fields of [psychology](#) and [learning](#) have also defined human intelligence:

Researcher	Quotation
<a href="#">Alfred Binet</a>	Judgment, otherwise called good sense, practical sense, initiative, the faculty of adapting one's self to circumstances...auto-critique.[4]
<a href="#">David Wechsler</a>	The aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with his environment.[5]
<a href="#">Cyril Burt</a>	Innate general cognitive ability[6]
<a href="#">Howard Gardner</a>	To my mind, a human intellectual competence must entail a set of skills of <a href="#">problem solving</a> —enabling the individual to resolve genuine problems or difficulties that he or she encounters and, when appropriate, to create an effective product—and must also entail the potential for finding or creating problems—and thereby laying the groundwork for the acquisition of new knowledge.[7]
<a href="#">Linda Gottfredson</a>	The ability to deal with cognitive complexity[8]
<a href="#">Sternberg &amp; Salter</a>	<a href="#">Goal-directed</a> adaptive behavior[9]

Essentially the same idea as Hutter's, but coming at it from a different angle and with different terminology, was put forward independently by Warren D. Smith in 2006 (see [13]).

Mathematical definitions have, as one advantage, that they could be applied to nonhuman intelligences and in the absence of human testers. The Hutter/Smith picture has a number of interesting consequences such as the theorem that "universal" intelligences exist which can emulate any other... ; that there are ways of creating quantitative "intelligence tests" which should enable serving as an objective gauge of progress in [artificial intelligence](#).

Theoretical and academic definitions of intelligence may not be as useful in clinical and therapeutic applications. For example, the clinical situation presented by those with borderline intellectual and adaptive functioning requires comprehensive analysis of all diagnostic, testing, educational placement, and psychosocial factors. This has been addressed in both the 8th (2005) and 9th (2009) editions of *Kaplan & Sadock's Comprehensive Textbook of Psychiatry* by Yale child psychiatrist Frank John Ninivaggi. MD.

## Theories of intelligence

The most widely accepted theory of intelligence is based on [psychometrics](#) testing or [intelligence quotient](#) (IQ) tests. However, dissatisfaction with traditional IQ tests has led to the development of a number of alternative theories, all of which suggest that intelligence is the result of a number of independent abilities that uniquely contribute to human performance.

# 12.

[http://www.coli.uni-saarland.de/~hansu/what\\_is\\_cl.html](http://www.coli.uni-saarland.de/~hansu/what_is_cl.html)

## What is Computational Linguistics?

By Hans Uszkoreit

### What is Computational Linguistics?

“Computational Linguistics is an interdisciplinary field which borrows from Linguistics and Computer Sciences and mainly Artificial Intelligence. It is made of two components one of which is theoretical and the other, applied.”

### Theoretical Computational Linguistics:

Scientists in the field of theoretical computational linguistics deal with the formal theories about linguistics and cognitive sciences. They form the computer programmes, which based upon these theories can further improve the linguistic competence by computers. Hence, linguistic theories and cognitive psychology play an important role in simulating linguistic competence.

### “Applies Computational Linguistics”

Applied CL touches on the outcome of modelling human languages. “The methods, techniques, tools and applications in this area are often subsumed under the term **language engineering** or **(human) language technology**”. The current CL systems are far from what is expected but they are still very useful for humans. The goal is to create software that is able, at least to some extent, to know about human language. While the problem for computer use is a “communication problem”, the CL specialists work hard to resolve this at least partially so that the bridge between human and computers can be widened.

### “Friendly Software Should Listen and Speak”

The revolution yet to occur is when computers are transformed from machines to human partners. This is possible when computers are enabled to speak through different language interfaces such as German, English, French, etc. and interact with us. The power to interact, along with other modes of interactivity such as pointing and mouse use have the potential to turn many applications live and kicking in the human environment and life.

“Machines can also help people communicate with each other.”

One of the main goals of use of computers has been full transfer of content from one language to another. This in fact, has been an earlier goal for scientists compared to human-computer understanding. The scientists faced tremendous weaknesses and realized the difficulty of the task. However, the partially enabled software still helped many including the information seekers in the current age.

**“Language is the fabric of the web.”**

The Internet as we know it is a combination of ideas put forward through various channels of media. This includes text, image, sound, and movie. This structure brings a wealth of exciting challenges for the computational linguistics, because we know that all of this content is only browsable through proper use of language. Filtering, Search, Categorization, and Translation would be a few of the possibilities. Not only is language needed to browse through this information, it should also be multilingual. This opens up possibilities for corporate collaboration, e-commerce, and online studying.

**“Out discipline combines ambitious visions and realistic applications.”**

The aim is full understanding by computers but there is still much work needed to do such thing. The aims and purposes for doing so are different among scientists and interested groups. However as more and more models are sketched using computers, the underlying pieces of knowledge come forward and the hidden patterns are revealed. Therefore it is acceptable to say “even today's language technologies full of clever short cuts and shallow processing techniques can be turned into badly needed software products.”

## 13.

<http://www.dfki.de/%7Ehansu/LT.pdf>

### Language Technology

By Hans Uszkoreit

Speech recognition

Spoken language is recognized and transformed in into text as in dictation systems, into commands as in robot control systems, or into some other internal representation. □



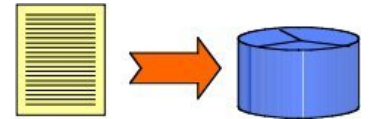
Speech synthesis

Utterances in spoken language are produced from text (text-to-speech systems) or from internal representations of words or sentences (concept-to-speech systems) □



Text categorization

This technology assigns texts to categories. Texts may belong to more than one category, categories may contain other categories. Filtering is a special case of categorization with just two categories. □



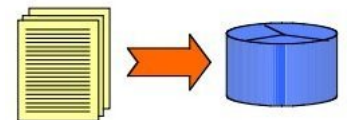
Text Summarization

The most relevant portions of a text are extracted as a summary. The task depends on the needed lengths of the summaries. Summarization is harder if the summary has to be specific to a certain query. □



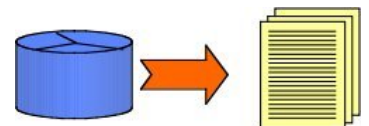
Text Indexing

As a precondition for document retrieval, texts are stored in an indexed database. Usually a text is indexed for all word forms or – after lemmatization – for all lemmas. Sometimes indexing is combined with categorization and summarization. □



Text Retrieval

Texts are retrieved from a database that best match a given query or document. The candidate documents are ordered with respect to their expected relevance.



Indexing, categorization, summarization and retrieval

are often subsumed under the term information retrieval. □

### Information Extraction

Relevant information pieces of information are discovered and marked for extraction. The extracted pieces can be:

the topic, named entities such as company, place or person names, simple relations such as prices, destinations, functions etc. or complex relations describing accidents, company mergers or football matches. □

### Data Fusion and Text Data Mining

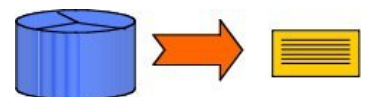
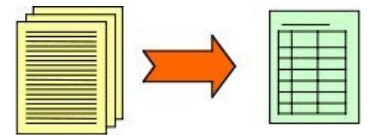
Extracted pieces of information from several sources are combined in one database. Previously undetected relationships may be discovered. □

### Question Answering

Natural language queries are used to access information in a database. The database may be a base of structured data or a repository of digital texts in which certain parts have been marked as potential answers. □

### Report Generation

A report in natural language is produced that describes the essential contents or changes of a database. The report can contain accumulated numbers, maxima, minima and the most drastic changes. □



### Spoken Dialogue Systems

The system can carry out a dialogue with a human user in which the user can solicit information or conduct purchases, reservations or other transactions. □



### Translation Technologies

Technologies that translate texts or assist human translators. Automatic translation is called machine translation. Translation memories use large amounts of texts together with existing translations for efficient look-up of possible translations for words, phrases and sentences. □



## 14.

<http://accurapid.com/journal/15mt.htm>

### Machine translation (MT), and the future of the translation industry

A usual concern among human translators is that machines will take over their business. On the one hand, prophets of doom announce a general crisis sending human translators onto the dole. On the other hand, any reasonable person who has tried MT software knows that human translation will be around for quite some time.

For how long? Yes, the output from any MT software is still laughable. But beware. MT software is still in infancy. And, given the pace of development in the computer industry (both in software and hardware), we may see, sooner than expected, an MT solution that provides decent translation. All it takes is a resourceful computer and better MT software. The hardware is here. The software will inevitable follow. Still, the output may not be good enough for public display, so the question turns into: will the future of human translation be... *proofreading computer output?*

The bad news is yes. It all boils down to how long it will be until computers produce decent translations. All they need is basically here. Neural Networks and Artificial Intelligence are slowly becoming better. The discoveries in other (well-funded) industries that are large

consumers of AI (civilian and military cybernetics, such as obstacle recognition, routing devices etc) will soon impact MT capabilities. Furthermore, the compilation of extensive knowledge bases such as dictionaries, glossaries, and translation memories will help improve machine translation.

As of going to press (November 2000), most of what MT software does is word-for-word translation followed by some grooming based on a set of rules. No surprise, the result is barely readable.

Let's take a comparison with what humans do with, for instance, calculation. There are actually 3 ways (and maybe more) of performing a calculation:

1. **computation.** When asked the result for  $145 + 133$ , we actually break down the operation into smaller ones, perform the necessary calculation and give the answer.
2. **memory.** When asked the result for  $8 \times 5$ , we immediately respond with recalling a table which we learned at school.
3. **common sense.** When asked whether  $1,450,000 \times 3,789$  is greater or smaller than 1, we give the gut answer "greater," although we do not actually perform the calculation (a computer will not respond as we do—it will calculate first, then give a final answer).

Kasparov can confirm that computers use methods 1 and 2, with considerable speed. We may say that method 3 is nice and poetic, but not efficient. This is a serious mistake, however. All serious IT engineers (there are thousands of them) that are concerned with what computers will do next are working precisely on that third method.

In other words, they are working precisely on how to turn you, a translator, into a proofreader. This may take a long time, but don't rejoice too fast. A long time, in the world of IT, is 3 to 5 years.

So-called fuzzy logic can make some people laugh. Those who were around in the microcosm of computer freaks of the eighties remember that fuzzy logic and fuzzy processors were regular topics in discussion groups and specialized magazines, but they were conspicuously absent in the nineties, as if those dreams had failed to deliver anything solid.

Fuzzy logic in itself is not difficult to implement. Any serious programmer can program fuzzy logic, or even better, implement a neural network. Once the fuzzy processor, be it soft or hard, has delivered a set of options to a given problem, the problem rests entirely on choosing the most "reasonable" option. If this applies to chess, the answer is pretty straightforward: reasonable means winning the game, period. In most human activities, however, and especially in language, the end purpose is not that simple. Consistency (human consistency) means that the answer has to match common sense, defined as the end result of countless learning situations which a person has lived since birth (we may distinguish one's personal trial-and-error situations, the wisdom acquired from education, plus the inborn instinctive knowledge: mature, nurture and nature). So the two-fold question is: can a computer memory store such a sum of knowledge? Can a program correctly process it and draw conclusions from it?

The answer to the first question, in the absolute sense, is no. The only way of *knowing* how ice cream tastes is to eat some. The only way of knowing how treason feels is to actually being betrayed, and so on. A computer can store a description of such things, but it cannot harbor feelings.

The answer to the second question, in the absolute sense, is no. Interpreting a knowledge base to which the program is fundamentally alien will inevitable lead to nonsense.

But if we stop dreaming of man re-creating man through science, and take our expectations to a reasonable level—can a machine actually perform some human tasks with reasonable accuracy, the answer is an obvious yes, and the time it will happen is soon. A decent MT machine is just around the corner. The computer *computes* chess, while Kasparov *plays* chess. A computer will never *understand*, but it can *translate*, at least to some extent. And, since translation without understanding is meaningless, the future of the human translator is proof-sensing what a machine has pre-translated.



## Part Four: Detailed and Field Specific Articles and Content

### 15.

<http://accurapid.com/journal/12xml.htm>

## XML and the Translator

By Alan K. Melby, Ph.D.

Why should you be interested in XML? Well, if you are not interested in HTML, then you probably won't be interested in XML. On the other hand, if you have translated a Web page or are even thinking about doing it someday, then you had better learn about XML.

As you know, a Web page is basically tags and text. Here is an extremely simple Web page:

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 3.2 Final//EN">
<HTML>
<HEAD>
<TITLE>ASTM Workshop Information</TITLE>
</HEAD>
<BODY BGCOLOR="WHITE">
<H1 ALIGN=CENTER>ASTM Workshop</H1>
<H2>Your Invitation to the March 2000 Workshop</H2>
<P>ASTM Subcommittee F15.48 on Language Translation cordially
invites you to actively participate in our committee work
to leverage common interests in the field of translation
and ensure premium quality service. We hope you will attend.</P>
</BODY>
</HTML>
```

Here are some of the tags:

1. <BODY>
2. <H1 ALIGN=CENTER>
3. </BODY>

Everything that is not a tag is text.

So far, so good. Now, what kinds of tags are there?

Tag number 1 (<BODY>) is called a start tag, since it comes at the beginning, and tag number 3 (</BODY>) is called an end tag, since it terminates the "element" which is everything between the start tag and its end tag. And, as in this case, the stuff in between can include other tags.

Tag number 2 is a start tag that has an "attribute" (ALIGN), and that attribute has a value (CENTER).

Another important concept is an "empty" tag. One empty tag is <BR>, which forces a line break. It is empty because it has no end tag.

Now what if we deleted the </P> tag in the sample Web page? Everything would still work, since the end tag is optional for a <P>, but would that make <P> an empty tag? Not at all! It just makes the end tag implicit. There is no implicit end tag for <BR>.

Now we have some terms under our belt (start tag, end tag, empty tag, implicit end tag, and attribute), and we can go on to XML.

HTML and XML are both members of the SGML family, but HTML is an "application" of SGML while XML is a "subset" of SGML. What is the difference? A big one. First of all, don't worry about exactly what SGML is. It is an ISO standard (number 8879), but that doesn't help you understand it. I have studied SGML for over ten years now, and I don't know all the ins and outs of it. It is a very abstract system for defining systems of tags. SGML is used to express the fact that <BODY> is an HTML tag, but <SHOULDER> and <EAR> are not. Someone (Tim Berners-Lee, to be exact) used SGML to define an application called HTML, and that changed the world because it was a key element in bringing the Internet to the average person. So it is important to make a good selection of tag names and how they fit together.

HTML has a fixed list of tag names. You can't just start using a new tag name, like <SUITCASE> and expect a browser to understand it.

The single biggest difference between HTML and XML, the difference that hits you in the face, is that in XML there is *no fixed list of tag names!* Instead, XML is a simplified form of SGML, and you can use it to define your own XML application that has its own tag names. In an XML application there could very well be a <SHOULDER> tag or even a <SUITCASE> tag. The list of tag names, along with other technical information about the tags, such as what attributes they can have and how they fit together, is found in the an awful looking thing called a DTD or a schema. May you never have to touch or even see a DTD or schema. Just the other day, I spent over twelve hours working a two persnickity DTDs. Just punishment for being ornery, some would say.

At at rate, now we can get the to the meat of the whole question of XML. Why would a translator even want to know about XML? I hinted at the answer at the beginning, when I said that those who translate Web pages will need to know XML. The reason is that XML is being combined with HTML in sophisticated Web pages already, and this trend should continue rapidly.

There are several ways to combine HTML and XML. One way is insert XML inside an HTML page. This is called an XML island.

Another way is to use XML *instead* of HTML for a Web page. This can be done already in

Internet Explorer 5. The XML describes the content and structure of the information to be presented; then, in order to display the Web page nicely, you attach to the Web page a set of instructions called an XSL stylesheet. An XSL stylesheet is kind of like a cascading style sheet in HTML 4, but much more powerful. It is actually a computer program written in XSL, which is a programming language designed specifically for use with XML.

So now you might be wondering how to deal with an XML page if you are asked to translate it.

Well, first of all, you will probably not translate the tag names. Just as you would not translate the tag names in an HTML page. This should be verified with whoever is requesting the translation.

Secondly, you need to get a list of the tag names for the particular XML application you are working with, so you will be able to spot typos and data corruption. Every XML document for that application must select from those tag names, unless they use what is called an "open" model, which allows creative tag names that cannot be found even in the DTD/schema. This is kind of wild, but you might run across it. Another exception is what is called a namespace, an advanced topic of another tutorial. You can spot a namespace by the prefix followed by a colon and a tag name.

Thirdly, there are a few key differences between XML and HTML you need to keep in mind, besides the basic difference (fixed set of tag names vs. different set of tag names for each XML application):

- a. Every empty tag has a slash at the end.

This will look strange at first. In the case of HTML, a break would be `<BR/>` to show that there is no end tag to look for.

- b. There are no implicit end tags.

There will never be a `<P>` with no `</P>`, for example. This is a hard and fast rule for XML, and sets it apart from HTML in a big way. Of course, things change. There is a kind of HTML that is XML compliant, and it, of course, has no implicit end tags. This is called XHTML, and is not yet very well known.

- c. Quotes are required on attribute values.

This is an easy one to mess up on. In HTML, you can leave off the quotes in some cases, such as `ALIGN=CENTER`, but in XML, it must be `ALIGN="CENTER"` or `ALIGN='CENTER'`.

There is much, much more to XML, and new things are happening in the XML world every day, but the few principles just explained should help a translator deal with the differences between HTML and XML and be able to ask questions instead of being completely lost in XML land

# 16.

<http://translationdirectory.com/articles/article1829.php>

## Open source translation management system

By ClientSide News Magazine

Interview with Gary Prioste

VP of Technology Solutions, Welocalize

Published - November 2008

**Q:** Please explain to our readers what the GlobalSight Open Source Initiative is? It is an Initiative that aims to drive the development of GlobalSight Ambassador, an industry-leading Translation Management System (TMS), through open collaboration. GlobalSight is a non-captive, vendor-independent community where participants are free to discuss, discover and build upon a TMS that can be shared by all.

GlobalSight was the name of the company that first developed the Ambassador product in 1997. For the next eight years, GlobalSight grew Ambassador from a software tool to develop and maintain multilingual websites, to the industry's first TMS that could automate the translation process and leverage previously translated material. In 2005 Transware acquired GlobalSight and continued to develop the product. In May of this year, Welocalize acquired Transware and subsequently inherited Ambassador.

GlobalSight is the name of the open source initiative, and Ambassador is the name of the TMS product.

...

**Q:** Tell us a little about the technical architecture of Ambassador.

**A:** It's a Java application, and will have a MySQL database. It will run in both a Linux and Windows environment. Further technical details will follow soon.

**Q:** Will the open source product be a SaaS, or an enterprise application?

**A:** Both. We plan on hosting an open source SaaS solution that the community can contribute to, and allowing users to download an enterprise version that they can run behind their firewall.

In both cases, the community can extend and enhance the product through the web-services API, or by modifying the core application.

**Q:** How will this open source project differentiate itself from others that have failed?

**A:** The idea of knowledge sharing and crowd sourcing is not new. We have seen several of these initiatives in our industry, from TAUS and TDA in translation automation, to TinyTM, OmegaT and TMOSS in translation memory. These projects and initiatives have gained traction only through collaboration and reciprocity. What about other open source projects that haven't succeeded? Perhaps it was due to a lack of initial investment, poor planning, bad timing, or lack of enthusiasm—most likely a combination of the above.

Whatever the case may be, I don't believe these projects were aligned with the true needs of the industry, at the enterprise level. And they certainly did not anticipate a change in those needs. Without support, advocacy and collaboration from users with a vested interest in developing and adopting a product that can achieve their business goals, the product cannot thrive in an open source community. We are listening carefully to what clients are saying. We want the initiative and product to succeed. And we believe the entire industry will benefit if it does.

## 17.

<http://www.xml-intl.com/web/guest/home>

## XML-Intl.

Better translation technology - XTM Suite 5.0 is released!

- Do you have large translation projects with multiple target languages?
- Do you need a shared server based Translation Memory with shared Terminology?
- Alternatively do you process many files and projects that involve many people?

XTM makes these projects easier to handle and control by automating many time consuming processes.

XTM automates:

- File processing
- Translation Memory Processes

...

Professional translation requires the translation buyer to interact with a translation

supplier. XTM Suite 5.0 is a set of web based translation tools that aids these interactions and gives you much more.

XTM is the only all in one translation tool which connects you to your suppliers and customers with the web browser.

We are XML-INTL. Our mission is to assist people in the business of translation to increase productivity, reduce costs and decrease through-put times through simple to use, yet powerful web technology.

XTM is currently the leading web 2.0 complete translation environment. XTM is powerful but simple to use: project managers can easily create, monitor and manage projects while translators, reviewers and terminologists can utilise the powerful fully-featured translation environment. XTM is an efficient collaborative environment where large projects can be divided between resources, who can then work together in real time, while benefitting from each other's translation memory. Based on Open Standards, XTM offers a new and easy way for people in the translation business to work. XTM is entirely scalable, suiting all sizes of organisation and providing an elegant solution to often complex problems. Call us on 01753 480 460 to discuss how XTM can help you.

## 18.

<http://accurapid.com/journal/49meeting.htm>

### **Automatic Translation in Multilingual Electronic Meetings**

*by* Milam Aiken, Mina Park, Lakisha Simmons, and Tobin Lindblom

**Volume 13, No. 3 - July 2009**

#### **Abstract**

Electronic meetings, e.g., chat rooms and bulletin boards, can be more efficient and effective than traditional, oral discussions, but until only recently, online groups speaking many languages could not benefit from machine translation (MT). Although it is possible for linguists to provide translations for the group members as they read comments during a multilingual discussion, this is not feasible for large groups and many languages. As a solution, we propose a fully automated multilingual meeting system, and an example of its use in a meeting with comments typed in English translated to Dutch and Russian illustrates its potential to reduce many multinational communication barriers. [For full article see <http://accurapid.com/journal/49meeting.htm>]

#### **Introduction.**

In the past, oral meetings involving speakers of multiple languages required participants to adopt a common language, e.g. English, or use interpreters. In the former case, all participants might not be fluent in the non-native language and could be uncomfortable speaking it. In the latter case, human interpreters could be expensive and difficult to schedule.

...

In this paper, we describe six Web-based machine translation services that can be used to assist with the understanding of foreign text and seven electronic chat systems that provide translations between language pairs. Then, we introduce a new, locally developed multilingual electronic meeting system that provides automatic translation among 41 languages. Finally, in a test of comprehension accuracy, we rank the languages using five simple phrases.

### Web-based machine translation

Since the late-1990s with the introduction of *Babelfish* on the Web (Yang & Lange, 1998), free, online translators have been available for use on text, documents, and Web pages. Currently, there are at least six free services (shown in Table 1) that provide support for different numbers of language pairs (e.g., English to Spanish, French to Russian, Chinese to English, etc.).

**Table 1:** Free Web-based translation services

Service	URL	Underlying MT	Language Pairs
<i>Babelfish</i>	<a href="http://babelfish.yahoo.com">http://babelfish.yahoo.com</a>	Systran	38
<i>Freetranslation</i>	<a href="http://www.freetranslation.com">www.freetranslation.com</a>	SDL	19
<i>Google Translate</i>	<a href="http://translate.google.com">http://translate.google.com</a>	Google	1,640
<i>Online-translator</i>	<a href="http://www.online-translator.com">www.online-translator.com</a>	PROMT	24
<i>Reverso</i>	<a href="http://www.reverso.net">www.reverso.net</a>	Reverso	19
<i>Worldlingo</i>	<a href="http://www2.worldlingo.com">http://www2.worldlingo.com</a>	Worldlingo	225

Using these Web sites, chat room participants could translate foreign comments, but conducting these translations can be confusing to group members (Flanagan, 1997). Group members are not likely to put forth the effort when faced with many comments in different languages, and a meeting facilitator providing translations for the discussion using these systems will be overwhelmed with the task once the group size reaches 5 or 6 with more than 2 or 3 languages (O'Hagan & Ashworth, 2002). More staff members could be added to help with the translations, but coordination would be difficult as they lost track of which comments were translated and which were not. Instead, automated translation is needed in multilingual electronic meetings.

### Automated multilingual meetings

Despite the first multilingual application appearing in the early 1990s that automatically translated between English and Spanish in an electronic meeting (Aiken, et al., 1992; Aiken, et al., 1994), at least two United States patents were filed several years later which claimed to do essentially the same thing:

...

Subsequently, at least seven applications (shown in Table 2) were developed that provide automatic translation for instant messaging between pairs of individuals.

**Table 2:** Online chat systems with automatic translation

Application	URL	Languages
Amikai	<a href="http://www.riskworld.com/PressRel/2000/00q3/PR00a076.htm">http://www.riskworld.com/PressRel/2000/00q3/PR00a076.htm</a>	9
Annochat	<a href="http://www.langrid.org/association/pangaeasupport/indexe.html">http://www.langrid.org/association/pangaeasupport/indexe.html</a>	4
ChatTranslator	<a href="http://www.chattranslator.com/">http://www.chattranslator.com/</a>	7
Free2IM	<a href="http://openaimblog.aol.com/2008/05/06/instant-language-translation-with-free2im">http://openaimblog.aol.com/2008/05/06/instant-language-translation-with-free2im</a>	13
Hab.la Realtime Chat	<a href="http://www.programmableweb.com/mashup/hab.la-realtime-chat-translation">http://www.programmableweb.com/mashup/hab.la-realtime-chat-translation</a>	41
IBM Lotus Sametime	<a href="http://my.advisor.com/doc/07484">http://my.advisor.com/doc/07484</a> <a href="http://www-01.ibm.com/software/lotus/sametime/">http://www-01.ibm.com/software/lotus/sametime/</a>	7
MeGlobe	<a href="http://meglobe.com/">http://meglobe.com/</a>	15
WorldLingo Chat	<a href="http://www.worldlingo.com/en/products/chat_translator.html">http://www.worldlingo.com/en/products/chat_translator.html</a>	15

However, multilingual meetings usually involve more than two people, and they often use more than two languages. We believe there is no system available that can accommodate such a group by automatically translating among several languages at once, but there is a clear need for such an application.

...

## Conclusion

The multilingual meeting prototype described here can support large groups using up to 41 languages with translations provided automatically within a few seconds via a link with *Google Translate*. Early results indicate a high level of comprehension for many translated comments, and future research will investigate the accuracy of more complex sentence translations as well as how the prototype performs with other languages.

## 19.

<http://nlp.stanford.edu/links/statnlp.html>

## Statistical natural language processing and corpus-based computational linguistics: An annotated list of resources Stanford University – Natural Language Processing

- 🔍 [Tools: Machine Translation, POS Taggers, NP chunking, Sequence models, Parsers, Semantic Parsers/SRL, NER, Language models, Concordances, Summarization, Other](#)
- 🔍 [Corpora: Large collections, Particular languages, Treebanks, Discourse, WSD, Literature, Acquisition](#)
- 🔍 [SGML/XML](#)
- 🔍 [Dictionaries](#)
- 🔍 [Lexical/morphological resources](#)
- 🔍 [Courses, Syllabi, and other Educational Resources](#)
- 🔍 [Mailing lists](#)
- 🔍 [Other stuff on the Web: General, IR, IE/Wrappers, People, Societies](#)

## 20.

<http://www.springer.com/computer/artificial/book/978-3-540-25031-9>

## Book Review: An Introduction to Language Processing with Perl and Prolog

An Outline of Theories, Implementation, and Application with Special Consideration of English, French, and German

The areas of natural language processing and computational linguistics have continued to grow in recent years, driven by the demand to automatically process text and spoken data. With the processing power and techniques now available, research is scaling up from lab prototypes to real-world, proven applications.

This book teaches the principles of natural language processing, first covering linguistics issues such as encoding, entropy, and annotation schemes; defining words, tokens and parts of speech; and morphology. It then details the language-processing functions involved, including part-of-speech tagging using rules and stochastic techniques; using Prolog to write phase-structure grammars; parsing techniques and syntactic formalisms; semantics, predicate logic and lexical semantics; and analysis of discourse, and applications in dialog systems. The key feature of the book is the author's hands-on approach throughout, with extensive exercises, sample code in Prolog and Perl, and a detailed introduction to Prolog. The reader is supported with a companion website that contains teaching slides, programs, and additional material.

The book is suitable for researchers and students of natural language processing and computational linguistics

## 21.

<http://linguistlist.org/issues/14/14-1536.html>

### Book Review: Computational Linguistics: Hammond (2003) LINGUIST List 14.1536 Thu May 29 2003

Anne Mahoney, Tufts University

Programming for Linguists is an introduction to computer programming using the Perl language, aimed at people who work with language. Although Hammond seems to envision it as a self-study guide, it would probably work better as a course textbook. It is a generally sound introduction to the language and to the notion of programming a computer.

Perl is a particularly nice language for text processing because of its wealth of pattern-matching and string-handling constructs. It is easy in Perl to say, for example, "find all words that end in a

vowel" or "replace every occurrence of the word 'cat' with the word 'feline.'" In addition, Perl is easy for beginners because it is interpreted rather than compiled: one simply writes a program and runs it, without explicitly having to turn it into machine code. As Hammond points out (p. 2), Perl is moreover available, and free, for every type of computer system in current use. I therefore agree that Perl is a good starting point for a linguist with a computational problem.

Hammond's intended audience is "a naive reader who may know nothing about programming" (p. ix). The reader who already knows another programming language and wants to pick up Perl will be better served by Wall et al. (2000) Hammond's naive reader, however, is expected to understand how to install software, how to use a text editor as distinct from a word processor, and how files and directories work. Although Hammond gives basic instructions on how to invoke an editor, how to invoke the Perl interpreter, and how to display the text of a Perl program, he leaves the reader helpless if anything goes wrong. While the details of using a text editor really are beyond the scope of the book, especially if the reader could be using any of several computing platforms, it is often the case that someone who has never thought about programming before has also never had occasion to use a text editor, change the path (set of directories from which executable programs can automatically be found), or install anything that requires configuration or compilation. Although Hammond sensibly suggests that some of these are "delicate tasks" and "you should seek assistance before attempting them on your own if you've never done this before" (p. 7), it would be useful to provide more concrete information about where such assistance might be available. A college- or university-affiliated linguist may be able to ask the school's "academic technology" group. If no such resource is available, the reader will want a good book on the relevant operating system, perhaps one introducing system administration or development.

...

## 22.

<http://billposer.org/Linguistics/Computation/Resources.html>

# Computational Resources for Linguistic Research Introduction

This page lists computational tools for doing linguistics. There is of course some overlap, but the

emphasis is on using computation to do what ordinary linguists want to do, not on computational linguistics for its own sake.

The page emphasizes free software that runs on Unix systems. The emphasis is on Unix for several reasons. First, that's what I myself use. Second, in my opinion Unix is the environment of choice for this kind of work. The Unix philosophy of making it easy to connect one small tool to another is just right for linguistic research. Third, Unix is strongly represented in the free software world.

**Contents** [For all current links, check out:

<http://billposer.org/Linguistics/Computation/Resources.html>

1. [Character Encoding](#)
2. [Fonts, Rendering, and Printing](#)
3. [Input Methods and Keyboard Layout](#)
4. [Extracting Text from Impure Formats](#)
5. [Regular Expressions and Other Pattern Matching](#)
6. [Unix Tools](#)
7. [Syntax](#)
8. [Text Corpus Databases and Searching](#)
9. [Obtaining Data From Web Sites](#)
10. [Sources of Electronic Text](#)
11. [Lexicography and Dictionaries](#)
12. [Concordances and File Comparison](#)
13. [Historical Linguistics](#)
14. [Sociolinguistics](#)
15. [Phonetics](#)
16. [Math and Statistics](#)
17. [Semantics](#)
18. [Other Software](#)
19. [Programming Languages](#)
20. [Structured Markup Languages](#)
21. [Miscellaneous](#)



Part Five: Further Graduate Studies in Detail

## 23.

<http://www.isi.edu/natural-language/MSCompLing/>

### **Information Sciences Institute - Master's Program in Computational Linguistics - University of Southern California – Courses Offered**

#### Courses Offered

##### *Required Courses*

**LING 530 Syntax (3 units, Fall)** An intensive introduction to the principle and methods of grammatical analysis. The structure of sample English clauses will be discussed, and the essential formal notions and empirical results will be presented. In particular, the central concept of multiplicity of representation will be developed in its various forms. The concept is inseparable from that of lexical representation of rule. Properties of relations between concomitant representations, as well as a taxonomy of such relations, are discussed and exemplified. The central grammatical relations are formalized and illustrated. A range of central phenomena will be discussed and analyzed, with appeal to cross-linguistic evidence and considerations.

**LING 534 Logic and the Theory of Meaning (3 units, Spring)** An introduction to logic in preparation for advanced work in semantics and linguistic theory. The language of first-order logic, and introduction to truth-theoretic semantics for formal and natural languages. Compositionality. Predication, and reference. Syntax and semantics of quantification. Generalized quantifiers and their lexical properties. Proper Nouns, definite descriptions and descriptive anaphora.

**LING 585 Computational Linguistics (3 units, Fall)** This course presents an overview of computational systems that process natural language, in particular examining the role of linguistic knowledge and the procedures that implement it in working systems. Topics covered include speech recognition and generation, computational lexicography, morphological analysis, natural language parsing and its relation to syntactic theory, lexical classes and lexical semantics,

and computational ontology. The course includes both hands-on and research components and emphasizes the expanding role linguistics can play in this emerging field, as well as how computational tools and techniques can contribute to linguistic research and theory.

### ***Recommended Prerequisite Course***

**LING 500 Structure of Language (3 units, Fall)** Development of analytical skills in syntax and semantics, with major attention to language universals and language typologies and their relevance to theories of language acquisition.

### **Breadth Requirement Courses**

**LING 533 Language Universals and Typology (3 units)** Introduction to language universals and typology.

**LING 538 Selected Topics in Romance Syntax (3 units)** Overview of selected topics in Romance Syntax within a comparative perspective and their contribution towards the understanding of a general theory of grammar. Prerequisite: departmental approval.

**LING 539 Japanese/Korean Syntax and Theoretical Implications (3 units)** Critical discussion of selected papers and dissertations on Japanese/Korean syntax and consideration of their theoretical implications. Prerequisite: departmental approval.

**LING 548 Lexical Semantics (3 units)** The primary focus of this course is on how meaning is constituted within the linguistic unit of the word. Several proposals for Lexical Decomposition are examined with both theory-internal and cross-linguistic empirical evidence being considered. Lexicalization Patterns form a major part of this evidence (that is, different ways in which languages group meaning elements together to form words). Other important topics include the question of how the meaning of a word (especially verbs) is related to its syntactic properties (mapping hypotheses), which in turn brings in the topics of Thematic/Semantic Roles, Selectional features and Verb Classes.

**LING 576 Psycholinguistics (3 units, Fall)** Theories of acquisition; sentence and discourse processing; language and thought. Prerequisite: departmental approval.

### ***Relevant Courses in Other Departments***

**EE 519 Digital Speech Processing (3 units, Fall)** Graduate introductory course on speech processing and speech recognition.

**EE 599 Advanced Topics in Speech Recognition and Spoken Language Engineering (3 units, Spring)**

**CSCI 561a Artificial Intelligence (3 units, Fall or Spring)** Foundations of symbolic intelligent systems. Agents, search, problem solving, representation, reasoning and symbolic programming. Prerequisite: CSCI455x. Since this is in Computer Science, this course requires permission from

the instructors.

**CSCI 544 Natural Language Processing (3 units, Spring)** This course covers the basic techniques of processing human language by computer. These include morphological analysis, parsing, semantic interpretation, and generation. We examine symbolic algorithms as well as some ways of acquiring linguistic knowledge automatically through statistical analysis. Techniques are presented in the context of applications like machine translation, text summarization, and information retrieval. Pre or co-requisite: CSCI 561a. Since this is in Computer Science, this course requires permission from the instructors.

**CSCI 562 Empirical Methods in Natural Language (3 units, Fall)** Acquiring computer-tractable linguistic knowledge has always been a bottleneck in building automatic translation, speech recognizers, summarizers, grammar checkers, and information management systems. Some of this knowledge can now be statistically extracted from large texts. We will examine the state-of-the-art in statistical modeling, supervised training, and bootstrapping. The approach will be experimental; software tools will allow students to build their own applications and measure performance. Prerequisite: CSCI 561a. Since this is in Computer Science, this course requires permission from the instructors.

## 24.

<http://www.isi.edu/natural-language/MSCompLing/>

# University of Southern California – Master’s Program in Computational Linguistics

Application information

**Who should apply:**

- Graduates in linguistics or related fields, or college seniors, interested in seeking a career in computational linguistics.
- Professional linguists interested in computational methods.
- Students looking to build a research-oriented resume before entering industry or a PhD program.

**Admissions requirements:**

Admissions to the Master's Program in Computational Linguistics requires:

- a completed bachelor's degree (or equivalent) in linguistics, mathematics, or a related field from an accredited institution
- a GPA of 3.0 or higher
- satisfactory GRE test scores
- -TOEFL test, for international students
- the ability to program, with expertise in such computer languages as LISP, C++, PROLOG, PERL, or JAVA

- proficiency in basic linguistics (phonology, morphology, syntax) and expertise in linguistics data analysis.
- advanced knowledge or the equivalent of at least two years of study at the college level of a human language other than English.

Applicants also need letters of evaluation from at least three professors or co-workers and a one or two-page statement of purpose. A major or work experience in a closely related field, such as Linguistics, Computer Science, or a particular language, is helpful but not required.

Students interested in the program who have not fulfilled the linguistics requirements (basic phonology, morphology, and syntax, and linguistic data analysis) or the recommended language requirement are strongly encouraged to fulfill them before entering the program. To help strengthen students' backgrounds in these areas, it may be possible to take a prerequisite course in Computer Science (Introduction to Programming Systems Design) and one in Linguistics (Structure of Language). Students should be aware, however, that the breadth requirements or elective courses may require additional prerequisites

## 25.

<http://www.cs.utoronto.ca/compling/Courses/courses.html>

# Graduate Studies in Computational Linguistics – University of Toronto – Courses Offered

## Our Courses

**CSC401H** (undergraduate) and **CSC2511H** (graduate)

**Natural Language Computing** 26L, 13T (Winter)

Instructor: [Gerald Penn](#)

An introduction to techniques involving natural language and speech in applications such as information retrieval, extraction, and filtering; intelligent Web searching; spelling and grammar checking; speech recognition and synthesis; automatic text summarization; pseudo-understanding and generation of natural language; and multi-lingual systems including machine translation. Methods covered will include *N*-grams, POS-tagging, semantic distance metrics, indexing, on-line lexicons and thesauri, morphological analysis and parsing, text markup languages and document structure, corpora and collections of on-line documents, corpus analyses. Software tools employed will include Perl.

*Prerequisites:* CSC228, STA220/250/257; CSC340 is recommended.

*Note:* CSC485/2501 and CSC401/2511 may be taken in either order.

**CSC485H** (undergraduate) and **CSC2501H** (graduate)

**Computational Linguistics** 26L, 13T (Fall)

Instructor: [Graeme Hirst](#)

Computational linguistics and the understanding and generation of natural language by computer. Syntactic processing. Semantics and semantic interpretation. Pragmatics, pronouns, definite descriptions, discourse context. Machine translation.

*Prerequisite:* CSC324H or experience in Lisp or Prolog.

*Note:* CSC485/2501 and CSC401/2511 may be taken in either order.

*Recommended preparation:* Students are urged to consult the instructor before enrolling.

Suggested background includes substantial computing experience and a course in AI, such as CSC384H, or some aspect of linguistics. Students in linguistics programs should consult the instructor.

**CSC2517H** (graduate)

**Discrete Mathematical Models of Sentence Structure** 26L (Not offered every year)

Instructor: [Gerald Penn](#)

Typed feature logic; mildly context-sensitive languages; parallel context-free grammars; tree-adjoining grammars; combinatory categorial grammar; pre-group grammars; tree transducers and tree-walking transducers.

**CSC2518H** (graduate)

**Spoken Language Processing**

Instructor: [Gerald Penn](#)

An introduction to working with speech in natural language processing systems. Topics include: articulatory and acoustic phonetics, prosody and information structure, introduction to digital signal processing of speech, automated speech recognition, text-to-speech synthesis, language models, dialogue modeling and dialogue systems. CSC2511H/401H1 is recommended (but not required) as a prerequisite.

**CSC2519H** (graduate)

**Natural Language Semantics** 26L (Fall 2007)

Instructor: [Gerald Penn](#)

An introduction to the study of meaning, its formal representation, its derivation from natural language syntactic structures, and its combination through inference with knowledge about the world. Topics may include: introduction to philosophy of language, compositionality, categorial grammar, quantification and plurality, underspecification, lexical semantics, word-sense disambiguation, lexical choice and nuances of meaning, calculating semantic distance, semantic interpretation in natural language processing systems, and reasoning with the event calculus in natural language.

**CSC2520H** (graduate)

**The Computational Lexicon** 26L (Not offered every year)

Instructor: [Suzanne Stevenson](#)

A computational lexicon is a highly structured repository of the rich syntactic and semantic knowledge about individual words in a natural language processing system. Two key issues will be the focus of this seminar course: the representation of lexical information, and its automatic acquisition. Topics will include: the organization of meaning and syntax in the lexicon; the interface between lexical semantics and its syntactic realization; the predicate argument structure of verbs; corpus-based approaches to automatic lexical acquisition and semantic/syntactic annotation of words; linking of statistical models to linguistic models of lexical properties; unsupervised learning of lexical relations; resolution of lexical ambiguities in natural language processing. Research papers will primarily focus on relevant research in computational linguistics, but we will also discuss work in linguistic and cognitive models of the human lexicon, and ways in which the engineering and cognitive approaches can inform each other.

**CSC2540H** (graduate)

[Cognitive Linguistics](#) 26L (Not offered every year)

Instructor: [Suzanne Stevenson](#)

**CSC2528H** (graduate)

[Advanced Computational Linguistics](#) (Not offered every year)

Instructor: [Graeme Hirst](#)

A seminar-style course that continues CSC 485/2501, and assumes the material presented therein. The course takes several topics of current research interest in computational linguistics, and studies them in depth. It emphasizes the interdisciplinary nature of computational linguistics. The interests of the class will determine exactly which topics are chosen. Auditors are welcome. *Prerequisite:* CSC 401/2511 or 485/2501 or permission of instructor.



This document was prepared by searching through online academic and general sources as

indicated at the beginning of each article. The most up-to-date content can be found online through the links provided at the top of each of the 25 pieces. The links are also provided below. All links are up-to-date as of January 30<sup>th</sup>, 2010. For questions or comments regarding this document please contact Bamshad Lotfabadi at [bamshad@bamshad.com](mailto:bamshad@bamshad.com)

Article 1 - [http://www.nsf.gov/news/news\\_summ.jsp?cntn\\_id=114063](http://www.nsf.gov/news/news_summ.jsp?cntn_id=114063)

Article 2 - <http://newsinfo.iu.edu/news/page/normal/10567.html>

Article 3 - [http://en.wikipedia.org/wiki/Association\\_for\\_Computational\\_Linguistics](http://en.wikipedia.org/wiki/Association_for_Computational_Linguistics)

Article 4 -

<http://www.translationdirectory.com/articles/article2037.php><http://www.translationdirectory.com/articles/article2037.php>

Article 5 –

Article 6 – <http://www.compling.washington.edu/compling/>

Article 7 – <http://www.isi.edu/natural-language/MSCompLing/>

Article 8 – <http://www.cs.utoronto.ca/compling/index.html>

Article 9 – [http://en.wikipedia.org/wiki/Computational\\_linguistics](http://en.wikipedia.org/wiki/Computational_linguistics)

Article 10 – [http://en.wikipedia.org/wiki/Artificial\\_intelligence#Philosophy](http://en.wikipedia.org/wiki/Artificial_intelligence#Philosophy)

Article 11 – <http://en.wikipedia.org/wiki/Intelligence>

Article 12 – [http://www.coli.uni-saarland.de/~hansu/what\\_is\\_cl.html](http://www.coli.uni-saarland.de/~hansu/what_is_cl.html)

Article 13 – <http://www.dfki.de/%7Ehansu/LT.pdf>

Article 14 – <http://accurapid.com/journal/15mt.htm>

Article 15 – <http://accurapid.com/journal/12xml.htm>

Article 16 – <http://translationdirectory.com/articles/article1829.php>

Article 17 – <http://www.xml-intl.com/web/guest/home>

Article 18 – <http://accurapid.com/journal/49meeting.htm>

Article 19 – <http://nlp.stanford.edu/links/statnlp.html>

Article 20 – <http://www.springer.com/computer/artificial/book/978-3-540-25031-9>

Article 21 – <http://linguistlist.org/issues/14/14-1536.html>

Article 22 – <http://billposer.org/Linguistics/Computation/Resources.html>

Article 23 – <http://www.isi.edu/natural-language/MSCompLing/>

Article 24 – <http://www.isi.edu/natural-language/MSCompLing/>

Article 25 – <http://www.cs.utoronto.ca/compling/Courses/courses.html>