

# Research Statement – Ariadna Font Llitjós

*Fostering communication and interaction among people through technology.  
Improving technology through people's interaction.*

## 1. Modern Communication and Computational Linguistics

In this information era, where large amounts of information are available in electronic format, the problem becomes how to make relevant information accessible to the people that need it at the right time and in their native language. Computational Linguistics, with its more practical counterpart **Natural Language Processing (NLP)**, is the interdisciplinary field that combines Linguistics and computational models of human language to study and facilitate communication between humans speaking different languages, and between humans and machines. For instance, NLP applications allow computers to follow spoken directions and answer questions in situations such as driving.

Any computational model of verbal communication, first transforms speech into text (speech recognition), then analyzes the text into its underlying structure (syntactic parsing), and optionally assigns meaning to it (semantic interpretation). Then the reverse process takes place: the model generates all the relevant underlying structures from the meaning (syntactic generation) and decides which specific word form (or set of forms) conveys the meaning of each concept (morphological generation); the last step is to convert text into speech following some letter-to-sound rules (speech synthesis).

### 1.1. Machine Translation

If communication occurs between two different languages, then Machine Translation (MT) comes into play. Over the past few years, the availability of massive amounts of parallel texts has lead MT researchers to work under the assumption that even syntactic structure is dispensable and that word-to-word and phrase-to-phrase translations are enough in order to translate from one language to another . Examples of such direct methods are Example-Based MT systems and Statistical MT systems. Most of such systems incorporate very little linguistic knowledge and are based upon information extracted from automatic word-to-word and phrase-to-phrase alignment, which are often erroneous.

When facing **languages with limited resources**, however, such a brute force approach to MT will not work. For the fast five years, I have worked on fast and inexpensive development of Rule-Based MT systems for such languages. The AVENUE project exploits available resources in resource-rich languages, such as English or Spanish, to transfer information to resource-poor languages, such as Mapudungun, Quechua and Inupiaq. I worked with native speakers in Temuco (Chile) and Cusco (Peru) in developing useful resources for them and building NLP tools that allow us to ultimately build MT systems for those languages (Font Llitjós *et al.* 2005).

Working together with end users of the technology we are building, has intensified my interest to make technology directly relevant to people.

Although statistics has significantly advanced the field in MT through the use of pure statistical methods for languages with large amounts of parallel bilingual text, researchers have not yet achieved satisfactory results. Accuracy and fluency of MT output take a hit and reflect that no sensible linguistic knowledge is guiding the system.

My dissertation work addresses this issue and approaches the problem from a different angle. Starting from a linguistically informed system, I explore ways in which bilingual speaker corrections augment and refine the grammar and the lexicon, thus achieving higher MT accuracy and fluency.

With increasing amounts of comparable online text in several languages, the idea of unsupervised learning from existing unannotated data (without requiring large amounts of expensive human labor) remains an appealing one (Meng and Siu 2002, Smith and Eisner 2004 and 2005, Turian *et al.* 2006). However, more recent trends in the field of Statistical MT have started to look at ways to incorporate at least syntactic knowledge into their approaches, so that the output of their MT systems is more grammatical and resembles more what humans would produce (Yamada and Knight 2001, Zollmann and Venugopal 2006, de Gispert and Mariño, 2006). I strongly believe that work in the other research direction is also necessary and can provide insights about the appropriate level of hybridization. In Font Llitjós and Vogel (2007), we explore adding statistical components to a Rule-Based MT system.

### ***Using Human Computation to improve Computational Systems***

#### **– Online Elicitation of MT output Corrections** (PhD thesis)

One of the most novel and interesting aspects of my thesis work is the idea of capitalizing on user time and availability in order to obtain annotated data through the Web, in this case corrected MT output, which can then be used to improve existing systems. The Translation Correction Tool (TCTool) is an online user-friendly interface that allows **non-expert, bilingual speakers** to minimally modify MT output in order to obtain a correct translation.

Given both the source language sentence and the target language sentence as well as the word alignments between them generated by the system, users can add, delete or edit words as well as move words around by dragging and dropping them. Alignments between words can also be deleted and added using the TCTool. User studies on English to Spanish MT output showed that users can correctly detect MT errors 89% of the time using the TCTool (Font Llitjós and Carbonell, 2004).

The idea of letting humans do what computers cannot yet do has also been exploited and formalized for image labeling (von Ahn and Dabbish, 2004).

## – **Automatic Improvement of MT systems** (PhD thesis)

Researchers have explored a variety of methods to include user feedback in the MT process. However, most MT systems have failed to incorporate post-editing efforts beyond the addition of corrected translations to the parallel training data for Statistical and Example-Based systems or to a translation memory database. My dissertation research centers on developing a largely automated approach that uses online post-editing feedback from non-experts to refine translation rules. Precise error correction information that is relevant to the system allows the **Automatic Rule Refiner** to trace the errors back to incorrect lexical and grammar rules responsible for the errors and to propose concrete fixes to such rules. For the most part, refinements involve adding morpho-syntactic information to existing rules.

Since this approach to improving MT output attacks the problem at its core, it generalizes beyond the input sentences corrected by bilingual speakers, and allows for correct translation of unseen data. Evaluation results on an English-to-Spanish Transfer-Based MT system show that by applying automatic refinements, higher translation accuracy can be achieved as measured by automatic evaluation metrics (Font Llitjós and Ridmann 2007, Font Llitjós *et al.* 2005).

## **2. Speech Synthesis** (Masters Thesis)

### – **Augmenting Pronunciation Models with Linguistic Origin**

Pronunciation of proper names that have different and varied language sources is an extremely hard task, even for humans. The main contribution of my Masters thesis was to develop a new approach to improve automatic pronunciation of proper names by modeling the way humans do it, and by trying to eliminate synthesis errors that humans would never make. My approach added information about different language and language family origin as features into the pronunciation models (**decision trees**). I also did exploratory investigations with **unsupervised clustering** of proper names to derive *language* classes in a data-driven way. With this approach, no language classes (Catalan, English, French, German, etc.) need to be determined a priori, but rather they are inferred from the names and their pronunciation. The clustering method took into account letter trigrams as well as their aligned pronunciation at training time (Font Llitjós 2001; Font Llitjós and Black, 2001).

Even though automatic evaluation measures did not detect a significant improvement on pronunciation accuracy, my work spurred interest on this line of research (Chung *et al.* 2003 and 2004; Kuo *et al.* 2005; Chen *et al.* 2006) and motivated me to develop an online site to collect human judgments on speech synthesis quality.

### – **Online Evaluation of Proper Name Pronunciation**

In 2002, I designed and implemented the US Pronunciation of Proper Names Site to evaluate and collect proper name pronunciations online. The internet proved to be a very successful medium both in terms of number of

evaluations and in terms of data collection. With an average traffic of 50 queries per day, I have been maintaining this site until recently. Information gathered is useful to improve pronunciation models, as well as to automatically correct the pronunciations in the CMU dictionary (Font Llitjós and Black, 2002).

### 3. Future Directions

In the next few years, I want to find more ways in which cutting-edge technologies can help people from different languages and cultures communicate and understand each other. I see the **internet as a great opportunity** to bring people together as well as to have easy access to user's time and knowledge (**human computation**) to help solve and improve NLP and Artificial Intelligence tasks that are still too hard for computers. I already started working on an online Translation Game that allows users to correct and validate MT output. If such a game is made available through a major web portal, this will result in a unique and extremely valuable collection of annotated data, which, at a large scale, can be used to improve not only Rule-Based MT systems, but also Statistical MT systems. Such an online game can also be very appealing to second language learners, who can test their skills by trying to correct MT output, and have the system score their performance and encourage them to try more sentences.

Current automatic **MT evaluation metrics** do not correlate well with human judgments. On the other hand, having monolingual speakers correct translations without direct access to the information contained in the original message, as proposed for the GALE project (Snover *et al.*, 2006; Sanders and Strassel, 2006) has severe limitations. My thesis work can be useful in designing a framework that **includes bilingual speakers' judgments on MT output**. Such a new approach to MT evaluation also has a clear advantage over commonly used reference translations, as it provides relevant corrections that address specific MT system errors. Therefore, MT systems are not punished for not having guessed which synonym or morpho-syntactic variation was used in independent human references.

During my dissertation research, I focused on rule adaptation as guided by a predetermined set of heuristics. However, when the underlying translation system is not accessible, it would be interesting to try to automatically learn a finite state transducer or a synchronous grammar that given an incorrect (corrupted) translation produces a correct translation. This could be done using **Machine Learning techniques** to process the data extracted with the Translation Correction Tool, namely negative examples (incorrect MT output) paired up with positive examples (corrected translation).

Finally, modern communication is **multimodal**, and finding ways to effectively integrate images, gestures, language (spoken and written) and context in order to provide better communication between humans and humans and machines is a very exciting field that will require interdisciplinary teams to work together.

## References

- Bennett, C., A. Font Llitjós, S. Shriver, A. Rudnicky and A. W. Black. 2002. "Building VoiceXML-based applications", *ICSLP 2002*. Denver, USA.
- Black, A. and A. Font Llitjós. 2002. "Unit Selection without a Phoneme Set". *IEEE Synthesis Workshop*. Santa Monica, USA.
- Chen, H.H., W.C. Lin, C. Yang, W.H. Lin. 2006. "Translating–transliterating named entities for multilingual information access". *Journal of the American Society for Information Science and Technology*. Volume 57, Issue 5.
- Chung, G., S Seneff and C Wang. 2003. "Automatic acquisition of names using speak and spell mode in spoken dialogue systems". In Proceedings of *NAACL–HLT*. Edmonton, Canada.
- Chung, G., C. Wang, S. Seneff, E. Filisko, M. Tang. 2004. "Combining Linguistic Knowledge and Acoustic Information in Automatic Pronunciation Lexicon Generation". *Interspeech-ICSLP*. Jeju Island, South Korea.
- de Gispert, A. and J. B. Mariño. 2006. "Linguistic tuple segmentation in ngram-based statistical machine translation". *9th International Conference on Spoken Language Processing (Interspeech'06)*. Pittsburgh, USA.
- Font Llitjós A. and S. Vogel. 2007. "A Walk on the Other Side: Adding Statistical Components to a Transfer-Based Translation System". To appear in *Syntax and Structure in Statistical Translation (SSST) Workshop at HLT-NAACL*, Rochester, USA.
- Font Llitjós A. and W. Ridmann. 2007. "The Inner Works of an Automatic Rule Refiner for Machine Translation". *METIS-II Workshop*. Leuven, Belgium.
- Font Llitjós A., J. Carbonell and A. Lavie. 2005. "A Framework for Interactive and Automatic Refinement of Transfer-based Machine Translation". *EAMT 10th Annual Conference*. Budapest, Hungary.
- Font Llitjós A., R. Aranovich and L. Levin. 2005. "Building Machine translation systems for indigenous languages". *Second Conference on the Indigenous Languages of Latin America (CILLA II)*, Texas, USA.
- Font Llitjós, A. and J. Carbonell. 2004. "The Translation Correction Tool: English-Spanish user studies". *LREC*, Lisbon, Portugal.
- Font Llitjós, A. and A. W. Black, 2002 "Evaluation and collection of proper name pronunciations online". *LREC*. Las Palmas, Spain.
- Font Llitjós, A. 2002. "Automatic Pronunciation of Proper Names using Language Origin Classes and Unsupervised Clustering". *ACL Student Session*. Philadelphia, USA.
- Font Llitjós, A. and A. W. Black. 2001. "Knowledge of Language Origin Improves Pronunciation Accuracy of Proper Names". *Eurospeech*. Aalborg, Denmark.
- Font Llitjós, A. 2001. "Improving Pronunciation Accuracy of Proper Names with Language Origin Classes". *Masters Thesis*. CMU-LTI-01-169. Carnegie Mellon University.
- Kuo, J.S. and Y.K. Yang. 2005. "Generating Paired Transliterated-cognates Using Multiple Pronunciation Characteristics from Web Corpora". Proceedings of the 18th *Pacific Asia Conference on Language, Information and Computation*. Tokyo, Japan.

- Meng H.H and K.C. Siu. 2002. "Semiautomatic Acquisition of Semantic Structures for Understanding Domain-Specific Natural Language Queries". *IEEE Transactions on Knowledge and Data Engineering*. Vol. 14, No. 1.
- Probst, K. 2005. "Automatically Induced Syntactic Transfer Rules for Machine Translation under a Very Limited Data Scenario". *Dissertation Thesis*. LTI. Carnegie Mellon University.
- Sanders G. and S. Strassel. 2006. "Post-Editing in the GALE Program". *Automated Post-Editing Workshop at AMTA*. Boston, USA.
- Smith N. and Jason E. 2005. "Guiding unsupervised grammar induction with contrastive estimation". In Working Notes of the *International Joint Conference on Artificial Intelligence Workshop on Grammatical Inference Applications*, Edinburgh, Scotland.
- Smith N. and Jason E. 2004. "Annealing techniques for unsupervised statistical language learning". In Proceedings of the *Annual Meeting of the Association for Computational Linguistics (ACL)*. Barcelona, Spain.
- Snover, M, B. Dorr, R. Schwartz, L. Micciulla and J. Makhoul. 2006. "A Study of Translation Error Rate with Targeted Human Annotation". *AMTA*. Boston, USA.
- Turian, J., B. Wellington, and I. D. Melamed. 2006. "Scalable Discriminative Learning for Natural Language Parsing and Translation". *20th Annual Conference on Neural Information Processing Systems (NIPS)*. Vancouver, Canada.
- von Ahn, Luis and Laura Dabbish. 2004. "Labeling Images with a Computer Game". *ACM CHI*. Vienna, Austria.