Soon after I moved to the United States in 2004, I became involved in a Russian lexicon project at Microsoft, and I had several contracts with the company through 2009. There, immersed in a mix of technology and language, I came to the conclusion that computational linguistics was where I belonged and that grammar engineering was something I could work on with enthusiasm. In Russia, languages were typically taught in a somewhat formal way, with more emphasis on grammar than on conversational skills, so I was already familiar with concepts of morphology, phonology, and syntax. I decided to pursue formal training in computational linguistics at University of Washington. Now that I am a first-year student in the UW Linguistics PhD program, I plan to explore the space between stochastic and rule-based approaches to building grammars of natural languages and eventually focus on finding efficient combinations of the two, providing modern currency in form of language technology to endangered and low-resource languages' communities as a **broader impact**.

Back in 2009 when I decided to apply to a computational linguistics program the biggest challenge on the way was that, having focused so much on languages and literature in my early years, I emerged with scarce knowledge of mathematics and undertrained quantitative reasoning abilities. To address this deficiency, I did a certain amount of self-study, got an associate degree from a local community college, and, as of June 2014, I have a Bachelor in CS from UW. Such rigorous preparation definitely serves as a solid foundation in my computational linguistics studies.

I started doing research the summer after my junior year joining two projects, one at the UW Electrical Engineering natural language processing lab (SSLI), and one at UW Linguistics. I spent two quarters working on the first one, which was more NLP-based, and three quarters on the second one, which is directly related to grammar engineering. This experience convinced me that I can and want to be in graduate school. Furthermore, it helped me realize that I prefer work on projects that are primarily concerned with problems in linguistics than language processing techniques alone.

The statistical machine translation project that I worked on for the EE department's Professor Katrin Kirchhoff was directed at developing an unsupervised approach to word sense disambiguation (Kirchhoff & Yang, 2012). Two separate metrics were used, one based on source - target word pair mutual information in a large corpus such as Wikipedia, and another on the relatedness scores of the word pairs obtained through the WordNet database. Because one of the most important clients of the machine translation industry is health care, the work was conducted on medical materials (English - Spanish pair), and the broader impact of these experiments will be an improvement in how fast health care professionals can process documents related to their patients. My responsibilities in this project included: corpus clean up and indexing; computing mutual information for all source word combinations based on the corpus index; building a graph based on Spanish WordNet; implementing efficient graph search; computing least-cost paths in that graph for all the source words that the graph covers. I completed these tasks while ensuring that the code that I design works fast on the 17GB Wikipedia corpus, on document lists that are over a million in length, and on numerous word combination possibilities. In general, machine translation and its potential interests me, and I especially appreciate that Professor Kirchhoff's approach is based not only on statistics, but on somewhat deeper processing as well, such as using the WordNet semantic information. While unsupervised approaches might indeed be the future, shallow language processing has its limitations, so a synthesis of the two might be the key.

The second project that I worked on and will be an RA for in Winter 2015, Dr.

Emily Bender's AGGREGATION, is concerned with automatic implementation of grammar fragments and is oriented towards documenting languages which are on the edge of extinction (G. M. W. C. J. Bender Emily M. & Xia, 2013). It is based on the HPSG, a formal theory of syntax (Pollard & Sag, 1994), and one of its goals is to provide field linguists with means to analyze their data faster and more effectively. An example of work in progress in that direction is adjusting a morphological analyzer which is part of the project, (Wax, 2014), to run on data from Matsigenka provided by field linguist Lev Michael, with a goal of inferring what the position classes are; I worked on this in winter and spring 2014. During summer 2013, I implemented a machine learning algorithm which classified small groups of languages based on which word order they exhibit. The ultimate goal of this in the whole project's context would be to train a classifier and then use it without supervision on large groups of languages for which there are no word order labels provided. The data vectors were produced by other members of the team before I joined, and I implemented a clustering classifier which increased accuracy on one of the labeled training sets, compared to a non-statistical approach. However, larger sets of training data proved to be not easily linearly separable, and my next goal in this project would be to find out if a non-linear approach will help. Generally, I would like to study what kind of functions are appropriate for different types of language data vectors at various stages, and how the parameters can be best tuned.

Before joining UW Linguistics PhD program, I took two graduate level courses from their computational linguistics program. One of them was dedicated to building a small grammar of Kolyma Yukaghir, a language of North-Eastern Russia, in a HPSG-based framework, with LinGO Grammar Matrix (E. M. Bender et al., 2002) serving as the starter toolkit. I appreciated the practicality of the HPSG formalism, and the idea of constraint unification seems to me to have great potential in terms of describing grammars in a parsimonious way. My work on the grammar led to an observation that one of the Case System options in the Matrix, which is also a typological resource, was attested only for one language family (Austronesian), while Dr. Bender and I decided it could provide an equally satisfying analysis of Kolyma Yukaghir, and that the option was therefore of a cross-linguistic nature. We then coauthored a paper which was presented at the HPSG Conference this year (Zamaraeva & Bender, 2014).

As teaching is something I would do throughout graduate school, I started working as teaching assistant in 2013. I picked an honors course that professor Anna Karlin was teaching. Named Brave New World, this course's objective was to attract diverse and talented freshmen to the UW CSE department. Simple programming tasks were explained and assigned, and the course as a whole was mostly dedicated to discussion of technology and its impact on society. We talked and blogged about massive open online education, robotics, social networks, and privacy. We tried to emphasize how ubiquitous computer science is and how important it is to study it, even if one's professional focus will ultimately be something else. As a TA, I helped to prepare some of the course labs, edited the pseudocode reference for students, and graded assignments, the grading rubric often being my responsibility as well. As a graduate student, I currently work as a teaching assistant for Dr. Bender's intro to HPSG syntactic theory course.

Finally, in 2013-2014 academic year I served as a vice-chair of Association for Computing Machinery - UW Women chapter. The organization's mission is in part to ensure moral support for women at the department and build a community in general. We organized useful events such as career workshops, social gatherings with faculty, dinner with industry affiliates, and participated in panel discussions about women in science. In addition, I personally tried to represent people who came to computer science from various backgrounds, not necessarily as strong numerically as their peers'. I believe that it would be beneficial for the computing community to recognize that certain shortcomings come not necessarily from lack of intellectual potential and most certainly not from gender, but rather from a circumstantial lack of experience which can still be gained, while different kinds of valuable experience that one did have a chance to develop can be shared, thus enhancing the field.

I demonstrated **intellectual** merit by graduating from one of the most competitive and rigorous computer science programs in the country (having come in with liberal arts background only), as well as by participating in two research projects for several quarters and coauthoring a paper while still an undergrad. The **broader impacts of my non-research activities** so far I consider increasing diversity in computer science through my work as a vice chair of ACM-W at UW and educating UW honors students about technology and computing as a teaching assistant . The **broader impacts of my research career** will have to do with making computational approaches accessible to everyone in the field of linguistics and beyond, particularly in documenting languages and analyzing their grammatical properties. Currently one needs to be familiar with particular formalism and often needs to be able to program in order to implement a grammar. I want to develop a system which only requires linguistic knowledge on one end and outputs a working grammar with minimal technical human involvement. As a **broader impact within the field of computational linguistics**, one the goals is to bring the best programming practices and efficient approaches to developing maintainable and reusable code to the linguistics community and thus significantly increase the amount of knowledge that is actually passed on from researcher to researcher. My hope is to help move grammar engineering forward, so that we can build highly efficient and reliable translation systems suitable for health care, document many of the endangered languages, and, most importantly, better understand the nature of human language.

References

Bender, E. M., Flickinger, D., & Oepen, S. (2002). The Grammar Matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In J. Carroll, N. Oostdijk, & R. Sutcliffe (Eds.), *Proceedings of the workshop on grammar engineering and evaluation at the 19th international conference on computational linguistics* (pp. 8–14). Taipei, Taiwan.

Bender, G. M. W. C. J., Emily M., & Xia, F. (2013). Towards creating precision grammars from interlinear glossed text: Inferring large-scale typological properties. In *Proceedings of the ACL 2013 workshop on language technology for cultural heritage, social sciences and humanities.*

Kirchhoff, K., & Yang, M. (2012). Unsupervised translation disambiguation for cross-domain statistical machine translation. In *Proceedings of AMTA.*

Pollard, C., & Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar.* Chicago, IL and Stanford, CA: The University of Chicago Press and CSLI Publications.

Wax, D. (2014). *Automated grammar engineering for verbal morphology.* University of Washington.

Zamaraeva, O., & Bender, E. M. (2014). Focus Case outside of Austronesian: An analysis of Kolyma Yukaghir. In S. Müller (Ed.), *The proceedings of the 21st international* *conference on Head-Driven Phrase Structure Grammar* (p. 1-20). Buffalo, NY.