The First Graduation Planning Meeting
October 12, 2018

1 Introduction

In order to graduate, you need to fulfill all the graduation requirements and apply for graduation. Important websites are listed at http://depts.washington.edu/uwcl/clms/. Some of them are explained below:

Requesting a MS degree: https://apps.grad.uw.edu/student/mastapp.aspx

Graduation requirements: http://depts.washington.edu/uwcl/clms/grad_req.html

Internship option :

- Internship requirements and deadlines:
  http://depts.washington.edu/uwcl/clms/internship_option.html
- Internship lists:
  https://canvas.uw.edu/files/50704960/download?download_frd=1 and
- Job search database:
  https://cldb.ling.washington.edu/livesearch-job-form.php

Thesis option :

- Thesis requirements and deadlines:
  http://depts.washington.edu/uwcl/clms/thesis_option.html
- Style manual for thesis:
  http://grad.uw.edu/for-students-and-post-docs/thesisdissertation/formatting-guidelines/
- Latex style file for thesis: (created by Fox)
  http://staff.washington.edu/fox/tex/uwthesis.shtml
- UW Libraries Linguistics thesis page:
  https://digital.lib.washington.edu/researchworks/handle/1773/4936
- Word templates used by previous CLMS students:
  Under /dropbox/CLMS_Thesis/Word_style_file on Patas.
- Completed CLMS theses:
  Under /dropbox/CLMS_Thesis/theses/ on Patas.

Project option :

- Project requirements and deadlines:
  http://depts.washington.edu/uwcl/clms/project_option.html

2 Graduation requirements

2.1 Courses

Students are required to take a minimum of **43** credits, as follows:

Required courses :

- Ling 550: Introduction to Linguistic Phonetics
- Ling 566: Introduction to Syntax for Computational Linguistics
- Ling 570: Shallow Processing for NLP
- Ling 571: Deep Processing for NLP
- Ling 572: Advanced Statistical NLP
- Ling 573: Systems/Applications

Electives :

- One linguistics elective
- One compling elective
- One more elective in compling or a related field

Internship/project/thesis : 10 credits of Ling 600 (internship/project) or Ling 700 (thesis option). If you are not sure which option you will eventually take, it is better to take Ling 700 as the graduate school requests a minimum number of 700 credits for a MS thesis.

If you waive a 4-or-5-credit required course (e.g., Ling 550) and replace it with a 3-credit elective, you need to take additional courses or additional Ling 600/700 credits to meet the 43-credit-minimum requirement.

2.2 Grades

Students must earn an average of at least 3.3 in each of the following groups:

- 550, 566, ling elective
- 570, 571, 572
- 573, compling elective, compling/related field elective

Any student failing to make a 3.3 average in any given group may either:

- Take an additional course (linguistics elective for Group 1, compling elective for Group 2 or 3)
- Retake a course

Any student earning less than 2.7 in a **required** course (550, 566, 570, 571, 572, 573) must retake the course. A grade below 2.7 in an **elective** may be made up with a different elective.

3 The Internship Option

Students who take the internship option should register for 10 credits of Ling 600 in the summer when they work as an intern.

3.1 Requirements

(A) Internship topic: The internship must be relevant to computational linguistics or human language technology more broadly. The internship (or portion thereof relevant to compling) must include at least 200 hours within a four-month period.

(B) Career office visit: In preparation for applying for internships, the student is encouraged to visit the Career Services Center for advice on developing a resume and related materials.

(C) Pre-internship report: The student must develop a pre-internship proposal (8 pages), in two steps:

1. Prior to applying for internship, a statement of the area of interest and proposed contributions, a discussion of why the company targeted is a relevant place to do this work, and a list of relevant references. (2-3 pages)
2. Once the student has been offered and accepted an internship, a literature review (based on the references identified above). (5 pages)

Students applying to multiple companies might develop 2-3 documents for Part 1, but only extend one to Part 2 for the company they end up working for.

(D) Self-Evaluation: At or near the end of the internship (prior to graduation), the student will write a self-evaluation which s/he will present to the internship supervisor for approval and then to the faculty advisor.

(E) Supervisor evaluation: In addition, we must receive a confidential written evaluation from the intern's supervisor, which references the self-evaluation. If this evaluation does not indicate satisfactory work, the internship will not count. The evaluation should be submitted through the Google Form available at: https://goo.gl/forms/5XtReX2l7tuVUqEs2

(F) Post-internship report: At or near the end of the internship (prior to graduation), the student must write a post-internship report (of 12-15 pages), with the following structure:

1. A description of the activities undertaken during the internship and their results (5 pages).
2. A discussion of how the CLMS course work related to/prepared the student for the internship work (3 pages).
3. A second version of the literature review. We expect this to be different from the initial version in incorporating the additional perspective gained in the course of the internship as well as any additional key papers that the student discovered in the course of internship work. (4-8 pages)

Students who are already employed in full-time (i.e., not internship) positions doing work relevant to computational linguistics or human language technology may also apply their activities at work towards the MS degree. Students in this position should negotiate with the CLMS faculty about how to adapt the above guidelines to their situation. For example, a student may propose a new project to undertake at work which applies what they've learned in the CLMS program.

All the documents will be reviewed by one or two CLMS faculty members, and if required by the internship companies, the faculty can sign non-disclosure agreements.

3.2 Applying for internship

Here is some information about potential internship opportunites. Please keep the CLMS faculty in the loop.

- Internship list:
  https://canvas.uw.edu/files/50704960/download?download_frd=1
- Job postings:
  https://cldb.ling.washington.edu/livesearch-job-form.php
- Previous internship positions:
  Microsoft (10+), Tegic Communication (6), Nuance/Voicebox Technologies (10+), Google (3), Fred Hutchinson Cancer Research Center (7), PNNL (2), PARC (2), Utilika (2), IBM (2), Group Health (2), Adapx (1), Amazon (7), Apple (1), Atigeo (2), Bose (2), Cambia Health (2), Cisco (1), Cataphora (1), Consilient (1), Linguistadores (1), Kiha Inc (1), PriceWaterHouseCoopers (1), Positronic (1), Sensory (1), ...

3.3 Deadlines

Pre-internship proposals (Part 1): 1/15

Career Services visit complete (recommended): 1/31

Internship: starts from June

Pre-internship proposals (Part 2: literature review): 6/15

Post-internship report (initial version): 7/20

Post-internship report (second version): 8/05

Internship self-evaluation with supervisor's approval: 8/05

Internship supervisor's evaluation: 8/10

Post-internship report (final version): 8/10

4 The thesis option

Students who take the thesis option should register for 10 credits of Ling 700. See Section 4.8 for detail.

4.1 Requirements

The MS theses will typically involve the implementation of working systems (or extensions or experimental evaluations thereof). In some cases, they may provide theoretical contributions instead. In contrast to MS projects, MS theses require a thorough literature review, are typically longer (30-50 pages), and represent the kind of research which could be presented at leading conferences in our field (ACL, COLING, NAACL, HLT, etc.). MS theses give experience with independent research as well as academic writing. The MS thesis is required for students who wish to petition for admission to the Department's PhD program, and could be beneficial to students who wish to apply to other PhD programs.

4.2 The common structure of a thesis

A thesis will typically involve the following chapters:

- Introduction
- Literature survey
- Methodology
- Algorithms, implementation, etc. (if appropriate)
- Experiments
- Discussion
- Conclusion and future work

You need to follow the university's formatting guidelines. A LaTeX style-file which is supposed to match the guidelines can be found at the urls mentioned in Section 1, but it is your responsibility to make sure that they are met.

4.3 Choosing a thesis topic

When deciding on your thesis project, it is important to keep three things in mind. See the Appendix for more discussion.

1. The project has to be small enough that it can be done in a few months' time. Quite often students come up with thesis topics that are far too broad. Once you've come up with a project idea, you'll likely want to immediately start narrowing the focus.

2. The thesis has to be novel in some way, such that it contributes to the knowledge of the field. Novelty can be broadly construed: You might be able to bring a new algorithm or approach to an old problem. Or maybe you'll have an idea that deals with a novel, unexplored data type. Or perhaps you want to do a comparative analysis of multiple algorithms on a well-defined task.

3. You need to find someone who can advise on the topic. If the theme of the project fits well with the skills and know-how of members of our program (Fei, Gina, Emily, and Ryan), you'll be all right. However, if you choose a project outside of our skill sets, it'll be your responsibility to locate a potential advisor, either someone in another department, or possibly someone in industry (perhaps a future internship advisor).

4.4 Finding an advisor

The CLMS faculty will work to let you know about our own on-going research (see Appendix C) and areas in which students can find MS project topics. In addition, we are happy to advise projects that are separate from our own research yet still within our expertise. If you would like to work on something outside our area of expertise, you are encouraged to approach other faculty on campus and/or send proposals to relevant groups in industry where you may do an internship (and we will help you make those connections).

Keep in mind that it is up to the faculty member or industry group to decide whether to take you on, and you'll need to sell your ideas to them. In addition, you'll want to get started on this early. Your must have your topic and advisor finalized by 3/1 if you plan to graduate at the end of the summer.

4.5 Finding a reader

In addition to finding an advisor, you also need to have a reader. Neither the advisor nor the reader needs to be a UW faculty member, but at least one of them needs to be CLMS faculty.

4.6 Deadlines

The following are the deadlines for Summer graduation.

late fall: Second graduation planning meeting: Faculty outline expectations for the first draft of thesis proposal.

1/3: 1st draft of the thesis proposal is due.

2/1: 2nd draft of the thesis proposal is due.

3/1: Final thesis proposal due. At this point, students are committed to pursuing this particular project.

4/1: Literature review due. Survey of strengths and weaknesses of previous approaches.

5/1: Methodology chapter due. How will you approach the problem? How will you evaluate the results?

6/1: Interim presentation of results. Prepare slides for presentation at a roundtable discussion with other students.

7/1: Results chapter due

7/7: First complete draft due

7/15: Feedback from advisor and reader on first draft

7/30: Second complete draft due

8/5: Feedback from advisor and reader on second draft

8/10: Final version of write up, ready for signature

In addition to the deadlines above, you must also meet the university filing deadlines for format checkup, etc.

If you plan to graduate in a quarter other than summer 2019, you should discuss this with your advisor, work out a new timeline, and get their approval in advance.

4.7 A few tips

It would be almost impossible to finish a thesis in three months; therefore, you need to start on your thesis as soon as possible and should finalize your thesis topic by March 1. In addition, you should consider the option of taking thesis credits earlier as explained above, and the possibility of double dipping.

Double dipping: Double dipping in this context means that you extend a project you choose for a course (e.g., Ling 575) and turn it into your thesis topic. Double dipping is not only allowed, but also encouraged; however, your thesis needs to be a significant extension over your course project, and you need to talk to your thesis advisor in advance to discuss the issue.

Starting early: You can register for 10 Ling 700 credits in the summer. However, if you want to spend more time on your thesis in your second or third quarters, you can register for the thesis credits in those quarters with your advisor's approval. For instance, you can take two courses and register for three thesis credits in the spring quarter, and take one course and register for the remaining seven thesis credits in the summer. Please be aware that most courses are not offered in the summer quarter, so plan in advance and talk to your advisor about your plan.

4.8 Policies with respect to thesis credits

Here are the general policies with respect to thesis credits:

- Full-time students are expected to enroll for all 10 thesis credits by the end of their fifth quarter of enrollment. Exceptions to this should be approved by the student's thesis advisor.
- Full-time students who fail to complete their MS theses within their fifth quarter and part-time students should determine the appropriate level of enrollment for their remaining terms in consultation with their their advisor. The number of credits should reflect the time and effort from the students and their advisors.

Students should consult with their advisor and reader about any specific requirements or expectations for the thesis.

5 Project option

The intent of the project option is for students to contribute to on-going research or related activities, not to carry out independent projects. Students interested in proposing their own projects should pursue the thesis option.

5.1 Supervisor and advisor

Under the project option, there are two possibilities:

- The student works with a third-party supervisor (a.k.a. "project supervisor"). The project supervisor does not need to be affiliated with UW. The choice of project and project supervisor must be pre-approved by one of the CL faculty, who serves as the student's advisor.
- The student works with a CLMS faculty member, who serves as both advisor and project supervisor.

In both cases, it is the student's responsibility to find a CompLing related project and someone on the project who is willing to serve as their project supervisor and allow them to work on the project. In addition, the project supervisor and the student's advisor must both agree on the content of the project (as discussed in the student's pre-project proposal) before the student starts working on the project.

5.2 Requirements

For a project to count towards the MS degree, all the following requirements must be met:

- Project topic: The project must be relevant to computational linguistics or human language technology more broadly. The project (or portion thereof relevant to compling) must include at least 200 hours within a four-month period.

	Thesis option	Internship option	Project option
Assigned an advisor?	No	Yes	Yes
Independent research?	Yes	No	No
Position in	university	industry	university
Normal length	6-9 months	3 months	3 months
Need to find	thesis topic advisor reader	Internship (supervisor)	Research Project (supervisor)
Main documents	thesis proposal thesis	pre-internship report post-internship report evaluation letters	pre-project report post-project report evaluation letters
Followup	publications	full-time job	(publications)

Table 1: Comparison of three MS options

- Pre-project report: the plan for the project, a discussion of why the project is related to CompLing, and a list of relevant references. The pre-project report needs to be approved by both the student's advisor and project supervisor.
- Evaluation letters: same as the ones in the internship option.
- Post-project report: same as the one in the internship option.

5.3 Deadlines

Pre-project proposals: 5/25

Post-project report (initial version): 7/20

Self-evaluation with supervisor's approval: 8/05

Supervisor's evaluation: 8/10

Post-project report (final version): 8/10

6 Choosing one of the three optionss

A comparison of the three options is shown in Table 1. In the past, about two thirds of students chose the internship option, and one fifth chose the thesis option. The project option was introduced in early 2010 and a number of students have taken it since then.

An internship is good for gaining work experience and making connections, and a summer internship often leads to a permanent position later. If your short-term goal is to find a CompLing job after graduation, we highly recommend the internship option.

The thesis option requires more work and has tighter deadlines. And it normally takes more than three months; 9-12 months is the norm. You need to plan your electives carefully in order to take advantage of double dipping. You also need to convince a faculty member to take you on. You should consider this option if you plan to continue on to a Ph.D. program in the near future.

The project option is for students who want to contribute to on-going research or related activities, not to carry out independent projects. It can also be a very good option for students with full-time positions which are not NLP-related. It can also serve as a backup plan for the other two options.

7 Final note

One final note: talk to your advisor about your plan for internship, thesis topic, course work, and the like. We are here to help you.

Also, it is very important for you to meet the timetable that you and your advisor have agreed on. You can revise the timetable but the revision must be approved by your advisor.

In the Appendix, we will discuss the thesis option further. It is divided into three sections. The first section details the three "dimensions" that should be part of any project: a task or application, an algorithm or algorithms, and a language, data type or domain that will be the focus of the project. The second section gives some ideas for projects that you might be able to pursue. The third section lists projects that CLMS faculty are working on. These projects were chosen specifically because there is a need for someone to work on them, they have good potential for thesis level work (novel ideas, clear evaluation, etc.), and because there are faculty in the program that can advise.

A The three dimensions of a thesis project

A thesis project can be characterized along three dimensions. In this section, we give some examples of each dimension. Note that the lists are far from complete. If you are interested in those areas, it is your responsibility to find a thesis chair who is an expert on those areas and who is willing to supervise you. But please talk to us and we will try our best to help you find a good match.

A.1 Dimension 1: the task

The first dimension is the task; that is, **what** is the problem you are trying to solve.

Basic NLP tasks: (some topics covered in LING 570 and 571)

- word segmentation
- morphological analysis
- POS tagging
- NP chunking
- Name entity tagging
- word sense disambiguation
- Parsing
- Grammar development: grammar induction, grammar engineering, grammar extraction.
- semantics: semantic analysis, semantic labeling.
- discourse: dialogue act, discourse structure, reference resolution.
- natural language generation
- dialogue

Applications: (some topics covered in LING 573)

- Machine translation
- Question and Answering
- Speech recognition
- Information extraction
- Sentiment detection
- Scientific paper classification
- Topic detection

A.2 Dimension 2: Algorithms

The second dimension is algorithm; that is, **how** are you going to solve the problem.

Symbolic approaches: Also called rule-based approaches

Corpus-based approaches: Machine learning methods (some covered in Ling 572)

- Supervised learning:
  - Decision Tree
  - Decision list
  - Transformation-based learning
  - Maximum entropy
  - SVM
  - Boosting
  - Bagging
  - Conditional random field
  - Neural networks
- Semi-supervised and unsupervised learning:
  - self-training
  - co-training
  - EM
  - Graphical models

Hybrid systems: combining symbolic and corpus-based methods

A.3 Dimension 3: data/language/domain

The third domension is about evaluation; that is, what data and evaluation metrics are you going to use to evaluate the effectiveness of your algorithm in solving the problem.

There are all kinds of data that can used: raw data, treebanks, dictionaries, WordNet, VerbNet, FrameNet, and so on. Data can be monolingual, bilingual, or multi-lingual. Data can be different domains: newswire, conversation, limited domain such as travel domain.

B Project ideas

For a typical project, you will choose elements from each of the three dimensions. Here are some project ideas. Once again, the list does not mean to be complete.

New algorithm: design a new algorithm for an existing task:
Ex1: write a parser with a totally new algorithm

New (task, algorithm) pair: Both the algorithm and task are well defined, but people have not tried to use the algorithm for the task before.

Ex2: CRF has been used for POS tagging. Now use CRF for parsing (see Sha and Fernadio's paper).

Adaptation (i.e., the data is new): people have used the algorithm for the task before, but not on this domain or for this language.

Ex3: Given a parser trained on WSJ data, try to improve its performance on the Brown corpus (cf. Mark Johnson's UW/MS talk)

Ex4: Given a parser trained on English data, retrain it on Chinese data, and try to improve its performance.

System combination: choose multiple algorithms from dimension 2, apply them to the same task in dimention 1, and combine the results so that the new system works better than each individual algorithm.

Ex5: Given Bikel, Collins, and Charniak's parsers, combine them to improve the performance.

Reranking: take an existing algorithm and generates topN candidates, and then use all kinds of features to re-rank the candidates.

Ex6: Use Charniak's parser to generate topN parses for each sentence, use all kinds of features to re-rank the list (see Mark Johnson's papers)

C Projects that the CLMS faculty are working on

Below is a short description of the projects that the CLMS faculty members are working on. If you are interested, please contact the faculty directly to discuss the potential of turning some subtasks into CLMS projects.

C.1 Emily's current research

If you are interested in working on something related to the Grammar Matrix or DELPH-IN please contact Emily. If you decide to work such projects, you should take Emily's Ling 567, offered this year in Spring quarter 2019.

C.1.1 Grammar Matrix

http://matrix.delph-in.net/

The Grammar Matrix is a system for facilitating the creation of DELPH-IN style precision grammars for typologically diverse languages. Every year, it gets one-three new libraries added through CLMS thesis projects.

Some open areas related to this project:

- Building a Matrix library for a linguistic phenomenon not yet handled, such as number names, complementation patterns beyond transitive/intransitive, imperatives, etc.

C.1.2 AGGREGATION

http://depts.washington.edu/uwcl/aggregation/

The AGGREGATION project is working on automatically answering the Grammar Matrix customization system's questionnaire on the basis of interlinear glossed text (IGT; typical examples in linguistics papers). That is, going straight from data as already annotated by linguists to operational

implemented grammars. In addition, we are working on looking at how this inference can shed light on patterns and also on inconsistencies in field linguists' data.

Some open areas related to this project:

- A project focusing on answering one part of the questionnaire from IGT.
- Contributing to on-going efforts to create useful UIs for exploring inference system outputs.

C.1.3 SEMEVAL and other Shared Tasks with DELPH-IN tools

http://erg.delph-in.net

If there is a shared task that you are interested in approaching for which you think the deep, precise representations provided by the ERG might be useful, that could also be a project it would make sense to work on with me.

C.1.4 Project (non-thesis) topics

I have several on-going projects that could use assistance and qualify for the project option:

- There was a major slow-down in ERG processing time as the grammar was developed for the 2018 release. Project idea: Hunting down the actual cause of this slow down. Ling 566 (next offered Winter 2019) + willingness to dive into large code base required.
- The Grammar Matrix customization system has a backlog of bugs that need squashing. Ling 567 required.

C.2 Fei's current research

Fei has worked on various subareas of CompLing. If you are interested in any of the topics below, please contact Fei. You should take ling570 in the fall quarter and ling572 in the winter quarter.

The RiPLes project: The goal of the RiPLes project is to bootstrap NLP tools for resource-poor languages (RPLs) with the information gathered from resource-rich languages. The seeds for bootstrapping come from interlinear glossed text (IGT) extracted from linguistic documents that are crawled from the Internet. Currently, we have collected about 200,000 IGT examples in close to one thousand languages, and the data is stored in the ODIN database. There are many subtasks in the project that could be turned into CLMS projects, ranging from linguistics-oriented to heavily statistical ones. Some examples are below:[1]

- Extending the ODIN database
- Building mini parallel treebanks for five to ten languages
- Creating language profiles

BioNLP projects

Fei is collaborating with Meliha Yetisgan and others at the UW Medical School on several bioNLP projects (e.g., phenotype detection, recommendation extraction, annotating medical events). More information is available at http://depts.washington.edu/bionlp/.

NLP Applications

[1] For some previous work on the project, please check the papers at http://faculty.washington.edu/fxia/research/papers_area.html.

Fei is interested in novel NLP applications, especially the ones related to social sciences (e.g., social psychology, political science). Given the large amount of social media data and the current political atmosphere, now is a good time to apply NLP techniques to analyzing social phenomena.

C.3 Ryan's Current Research

Ryan has a few projects related to data mining for resource poor languages, if you are interested in getting your hands dirty with some noisy, real-world data mining.

C.3.1 Data Mining/Low-resource Languages

- Building language models and reapplying OCR to bilingual research papers — application to improved acquisition of resource-poor language data, or historical documents of interest to those in the digital humanities.
- A project improving the IGT enrichment framework from INTENT (https://github.com/rgeorgi/intent) for INTENT2, using scikit-learn to replace off-the-shelf Java tools with pure python.
- Creating new evaluation data for the automatic enrichment techniques (***especially*** if you happen to have expertise in a non-Indo-European language!)

C.3.2 Distributional Semantics/Ethics & Abusive Language

- A replication project on detection of abusive language using word embeddings from a different dataset (provided)
- Project on annotating data for task of detecting consentful vs. coercive/violent language — (*warning that this task would require analyzing violent/potentially triggering content*)

C.4 Gina's current research

Gina has worked on a range of areas in speech and language processing. If you are interested in any of the topics or projects below, please contact Gina. If you are interested in speech-oriented projects, you should take Ling 550 or 553 (or equivalent).

EL-STEC: Shared Task Evaluation Challenges for Endangered Language Data The EL-STEC project that I'm working on with Emily aims to build and carry out Shared Tasks that develop systems that relieve bottlenecks in endangered language documentation, while pushing the state of the art in speech and language processing. Possible projects include:

- First-pass IGT creation: building baseline systems for morphological segmentation, morph glossing, and word glossing
- Language and speaker ID: building baseline systems for language and speaker identification, especially with small amounts of training data
- Genre classification: building baseline systems for classifying endangered language archive data genres, such as conversation, monologue, or chanting.
- Evaluation systems: Building systems to process archive data and evaluate task systems, including building virtual machines.

For tasks relating to first-pass IGT creation, you should talk to Emily. For other tasks, contact Gina.

NLP for low-resource languages This project involves resource and tool creation for natural language processing of low-resource languages, especially using techniques with leverage typologically similar languages or language universals.

Possible projects include:

- Cross-language projection of lexical resources
- Comparing cross-language projections by orthographic, phonetic, and/or typological similarity
- Tuning of machine translation systems for low-resource settings
- Implementing graph-based models for lexical resource augmentation

Other topics: In addition to these main projects, Gina is also interested in:

- Recognizing sentiment, subjectivity, and stance in speech
- Analyzing turn-taking in human-human and human-computer dialog
- Spoken dialogue systems, dialog state tracking, including evaluation
- Discourse analysis, including discourse parsing and anaphora resolution
- Topic segmentation