LING YYY Technology for Language Documentation and Revitalization
PROSPECTIVE SYLLABUS

Instructor:	Atticus Harrigan	**Time:**	TBD
Email:	galvin+lingYYY@ualberta.ca	**Place:**	TBD

Course Page: TBD

Office Hours: TBD

Required Texts: See reference list. All readings are either freely available online or available through the university library. Links to these resources will be posted on the class website.

Objectives: This course introduces students to the use of technology for the purposes of documenting and revitalizing languages. By the end of this course, students will be able to:

- Identify and describe the different types of technology used in language documentation and revitalization
- Recognize and respond to ethical issues of language technology
- Formulate and evaluate a research plan involving the use or creation of technology for language documentation or revitalization

Prerequisites: Undergraduate coursework in each of the following: morphology, phonology, and syntax. Students are *highly* encouraged to have taken a previous course in computing or programming.

Email: When emailing me about this course, please use the email galvin+lingYYY@ualberta.ca. I'm going to setting up a rule to place all emails sent there into a priority inbox. I plan to respond to all emails within 24 hours, but if galvin+lingYYY@ualberta.ca is not the email your question is sent to, I can't promise I'll be able to stick to that timeline.

Course Content

Lectures: Success in this course relies on students having read the assigned readings before class. Lectures will be discussing materials stemming from the readings. Although no attendance grade is given, regular attendance is essential for success in this course. Although this course will make use of a computer lab with all required software, You are encouraged to bring personal laptops to work on if available.

Marking & Grading: Marks for assignments will be given in percentages, and an overall course mark will be calculated according to the weights given below. Letter grades will be assigned to the overall mark according to the University guidelines. There is no curve for the final letter grade (i.e. the distribution of grades is not predetermined by a quota that requires a certain number or percentage of grades at a particular level). As per the University Calendar, marks and grades are based on a combination of absolute achievement and relative performance in a class. Grade cutoffs can thus differ between courses. Representative evaluative course material (such as practice/Test Yourself quizzes) will be available on the course website.

Course Outline:

Date	Topic	Readings
Week 1	Introduction to Language Documentation	[1] & [2]
Week 2	What is Language Technology?	[3] & [4] (ToC)
Week 3	Documenting Languages: ELAN, NLTK, and SIL tools	[5] & [6] & [7]
Week 4	Revitalization Technology: Morphological Models, I-CALL, and Speech Technology	[8] & [9] & [10] & [11]
Week 5	Building a Finite State Transducer	[12] & [13] & [14] (optional)
Week 6	Language Processing in Python with the NLTK	[15] (ch. 1 and 3)
Week 7	Language Processing in Python with the NLTK	[15] (ch. 3 and 11)
Week 8	Introduction to Neural Networks	[4] (ch. 7)
Week 9	Using NN frameworks: OpenNMT	[16]
Week 10	Neural Approaches to Documentation and Revitalization	[17] & [18] & [19]
Week 11	Project Presentations	

Evaluation Policies

Evaluation Metrics

Assignment	Weight	Date Due
Describing Language Technology	20%	End of Week 3
Using and Critiquing Documentary Software	20%	End of Week 4
Using and Critiquing Revitalization Software	20%	End of Week 6
Proposal for Final Project	10%	End of Week 8
Final Project Presentation	10%	Week 11
Final Project Write up	20%	Last Day of Class

Assignments: This course is assignment based. There will be no exams. Instead, there will be four assignments given throughout the semester. The first three assignments are each worth 20% of your grade, while the final project is worth 40% of your final grade (10% for the proposal, 10% for the final write up, and 20% for the project write up.

Each assignment will be submitted electronically to me through the class website, with the exception of the final project presentation, which will be given in person in the final week of classes. All assignments must be submitted as a PDF file. Submissions should be in 12 point font for body text and 10 point font for tables or captions. Documents should have 1 inch margins on all sides. Beyond this, there are no specific formatting requirements. Details for assignment lengths and contents will be given in assignment handouts uploaded to the course website. A *summary* of each assignment is given below.

Assignment 1: Describing Language Technology: This assignment will ask you to choose a piece of language technology for an under-resourced or endangered language. You will describe the underlying technology, identify the goal or problem the technology is attempting to remedy, and engage with how the technology does or does not engage with under-resourced endangered language community members.

Assignment 2: Using and Critiquing Documentary Software: In this assignment, you will make use of some documentation software (e.g. Saymore, ELAN, etc.) to record yourself saying a paragraph (roughly 150 words) in a language you are fluent in. You will also use transcribe this recording. Describe your experience in recording and transcribing, paying attention to the difficulties or ease you experienced in this process. If you have criticisms, suggest some ways these could be remedied.

Assignment 3: Using and Critiquing Revitalization Software: Similar to the previous assignment, you will pick some piece of revitalization software (morphological analyzers, I-CALL, or speech technologies) and attempt to make use of it. You do not need to document an under-resourced language, but must attempt to use the software to engage in an act of language revitalization (with any language, endangered or not) as defined in class. You will describe the process, difficulties, and ease with which you completed this task. You will highlight the accessibility of this software, and describe how it does or does not engage with endangered language communities. Criticisms should be accompanied by suggestions for improvements.

Assignment 4: Final Project Proposal: You will provide a brief proposal for the building of some software for language documentation or revitalization. You should be able to complete this final project within the course of the semester. This means the proposed project should be realistically ambitious. Projects do not need to create entirely new frameworks, and can instead be instantiations of existing technologies (e.g. creating a small FST model of a particular linguistic feature, training a speech synthesizer, creating/analyzing a corpus using NLTK). This proposal will graded on how achievable it is and how well it engages with technologies discussed in class. The proposal should describe the goal of the project, the implications of the technology on documentation or revitalization, and how the technology will engage and serve under-resourced language communities.

Assignment 5: Final Project Presentation: The Project presentation will be judged separately from the project itself. You will be graded on their presentation skills and ability to successfully communicate academic research.

Assignment 6: Final Project: The final research project will be graded *not* on the quality of their code, but on the write up of the process. You are expected to describe a problem or need in endangered or under-resourced language communities, identify and/or create some technology to address this need, evaluate or provide an evaluation framework for the technology, and describe the ethical and moral considerations of the technology.

Student Success

Life can be stressful. It is important to know that **you are not alone**. I want to make it very clear that your performance in this (or any) class is **not a reflection of your value as a person**. If you ever feel overwhelmed or stressed out by this class in particular, or your student life in general, please feel free to come and talk to me. If you find that weird, or you prefer other types of support, there are many resources available to you as a uAlberta student, such as the Peer Support Centre and Counselling Services. Please take care of yourself. Your grades in university are not more important than your health.

Other Required Information

Policy about course outlines can be found in Section 23.4(2) of the University Calendar.

Academic Integrity & Academic Honesty: The University of Alberta is committed to the highest standards of academic integrity and honesty. Students are expected to be familiar with these standards regarding academic honesty and to uphold the policies of the University in this respect. Students are particularly urged to familiarize themselves with the provisions of the Code of Student Behaviour and avoid any behaviour that could potentially result in suspicions of cheating, plagiarism, misrepresentation of facts and/or participation in an offence. Academic dishonesty is a serious offence and can result in suspension or expulsion from the University. All students should consult the information provided by the Vice-Provost and Dean of Students regarding cheating and plagiarism in particular and academic dishonesty in general. If in doubt about what is permitted in this class, ask the instructor. Students involved in language courses and translation courses should be aware that on-line "translation engines" produce very dubious and unreliable "translations." Students in language courses should be aware that, while seeking the advice of native or expert speakers is often helpful, excessive editorial and creative help in assignments is considered a form of cheating that violates the code of student conduct with dire consequences. An instructor or coordinator who is convinced that a student has handed in work that he or she could not possibly reproduce without outside assistance is obliged, out of consideration of fairness to other students, to report the case to the Associate Dean of the Faculty. See the Academic Discipline Process.

Learning and Working Environment: The Faculty of Arts is committed to ensuring that all students, faculty and staff are able to work and study in an environment that is safe and free from discrimination and harassment. It does not tolerate behaviour that undermines that environment. The department urges anyone who feels that this policy is being violated to:

- Discuss the matter with the person whose behaviour is causing concern; or
- If that discussion is unsatisfactory, or there is concern that direct discussion is inappropriate or threatening, discuss it with the Chair of the Department.

For additional advice or assistance regarding this policy you may contact the Office of the Student Ombuds. Information about the University of Alberta Discrimination and Harassment Policy and Procedures is described in UAPPOL.

Recording of Lectures: Audio, video, or photographic recording of lectures, labs, seminars or any other teaching environment by students is allowed only with the prior written consent of the instructor or as a part of an approved accommodation plan. Naturally, in an online class, much of the course material is already recorded; any recorded material is to be used solely for personal study, and is not to be used or distributed for any other purpose without prior written consent from the instructor. You are prohibited from re-distributing any course materials as described above; this includes uploading slides and other materials to websites such as CourseHero.

Attendance, Absences, and Missed Grade Components: Regular attendance is essential for optimal performance in any course. In cases of potentially excusable absences due to illness or domestic affliction, notify your instructor by e-mail within two days. Regarding absences that may be excusable and procedures for addressing course components missed as a result, consult the Calendar regarding Attendance and

Examinations sections of the University Calendar. Be aware that unexcused absences will result in partial or total loss of the grade for the “attendance and participation” component(s) of a course, as well as for any assignments that are not handed in or completed as a result. It is your responsibility to make up work and learn material for missed classes. In an online course, “attendance” begins to feel less relevant; however, be sure to read all communication around this course - including emails from the instructor, posts to the discussion forums, and documents posted on the course website. Recognize that those who choose not to read communication surrounding this course must assume whatever risks are involved.

Policy for Late Assignments: Except in extenuating circumstances, there are no redos for missed assignments. If you present a legitimate reason for missing an assignment, the weight for that assignment will be redistributed to the remaining (that is, future) grade items in the course. If you have concerns with the timing of any of the assignments, email me as soon as possible.

References

[1] Lenore A Grenoble and Lindsay J Whaley. *Saving languages: An introduction to language revitalization.* Cambridge University Press, 2005. 2

[2] Aidan Pine and Mark Turin. Language revitalization. In *Oxford Research Encyclopedia of Linguistics.* 2017. 2

[3] Sjur Moshagen, Tommi A Pirinen, and Trond Trosterud. Building an open-source development infrastructure for language technology projects. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 343–352, 2013. 2

[4] Dan Jurafsky and James H. Martin. Speech and Language Processing (3rd ed. draft), year = 2020, url = https://web.stanford.edu/ jurafsky/slp3/. 2

[5] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. ELAN: a professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559, 2006. 2

[6] Stuart Robinson, Greg Aumann, and Steven Bird. Managing fieldwork data with Toolbox and the Natural Language Toolkit. *Language Documentation & Conservation*, 1(1):44–57, 2007. 2

[7] Sarah Ruth Moeller. Review of SayMore, a tool for Language Documentation Productivity. *Language Documentation & Conservation*, 8:66–74, 2014. 2

[8] Ryan Johnson, Lene Antonsen, and Trond Trosterud. Using finite state transducers for making efficient reading comprehension dictionaries. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 59–71, 2013. 2

[9] Antti Arppe, Jordan Lachler, Trond Trosterud, Lene Antonsen, and Sjur N Moshagen. Basic language resource kits for endangered languages: A case study of plains cree. In *Proceedings of the 2nd Workshop on Collaboration and Computing for Under-Resourced Languages Workshop (CCURL 2016), Portorož, Slovenia*, pages 1–8, 2016. 2

[10] Nathan Thanyehténhas Brinklow, Patrick Littell, Delaney Lothian, Aidan Pine, and Heather Souter. Indigenous language technologies & language reclamation in canada. 2

[11] Candace Kaleimamoowahinekapu Galla. Indigenous language revitalization, promotion, and education: Function of digital technology. *Computer Assisted Language Learning*, 29(7):1137–1151, 2016. 2

[12] Mans Hulden. Foma: a finite-state compiler and library. In *Proceedings of the Demonstrations Session at EACL 2009*, pages 29–32, 2009. 2

[13] Atticus G Harrigan, Katherine Schmirler, Antti Arppe, Lene Antonsen, Trond Trosterud, and Arok Wolvengrey. Learning from the computational modelling of plains cree verbs. *Morphology*, 27(4):565–598, 2017. 2

[14] Sarah Beemer, Zak Boston, April Bukoski, Daniel Chen, Princess Dickens, Andrew Gerlach, Torin Hopkins, Parth Anand Jawale, Chris Koski, Akanksha Malhotra, et al. Linguist vs. machine: Rapid development of finite-state morphological grammars. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 162–170, 2020. 2

[15] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009. 2

[16] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July 2017. Association for Computational Linguistics. 2

[17] Mans Hulden, Markus Forsberg, and Malin Ahlberg. Semi-supervised learning of morphological paradigms and lexicons. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 569–578, 2014. 2

[18] Lane Schwartz, Emily Chen, Benjamin Hunt, and Sylvia L.R. Schreiner. Bootstrapping a Neural Morphological Analyzer for St. Lawrence Island Yupik from a finite-state transducer. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 87–96, Honolulu, February 2019. Association for Computational Linguistics. 2

[19] Robert Jimerson, Kruthika Simha, Ray Ptucha, and Emily Prud'hommeaux. Improving ASR output for endangered language documentation. In *The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, 2018. 2