Syllabus
GNED 495: Computational linguistics

Spring 2019

Time:	TR 2pm
Location:	Ruffner 350
Website:	`http://cs.longwood.edu/courses/gned495`

An investigation of the field of computational linguistics, integrating programming skills with an understanding of the syntax, semantics, and other aspects of English and other natural languages. Prerequisite: CMSC 140 (FQRC) with a grade of C− or better and a pillar-level World Languages course (FGLO). CMSC 160 can substitute for 140 for students who have already met their FQRC requirement. 3 credits. PQRC, WI.

Professor:	Don Blaheta
Office:	Ruffner 334
Phone:	x2191
Email:	`blahetadp@blahedo.org` or `blahetadp@longwood.edu`
Office hours:	Mon/Wed 1:30-2:30pm; Tue 11-noon; Fri 2-3pm

Overview

Careful use of languages, and understanding of how languages work, have themselves long been essential to participation in global life, and in civic life even in comparatively monolingual milieux (where they tend to go by names like "grammar and rhetoric"). The scientific study of language—that is, linguistics—has always been informed by empirical observation, and as digital computers became more powerful, computational linguistics per se has become an important pillar of the field. Yet outside the academic world, and even in some places inside it, people often present arguments, linguistic in nature, that are contradicted by the data. Or, that could be supported by data, but are not. Being able to identify such "truthy" arguments, and to combat them, is a skill vital to civic life, with further implications for information literacy as well. By integrating knowledge and skills from math, computer science, languages, and linguistics, students in this course will gain perspective on such arguments and learn to make their own data-driven arguments and decisions.

Textbook and resources

The textbook is *Language and computers*, by Dickinson, Brew, and Meurers (ISBN 978-1-4051-8305-5). There will be regular in-class and out-of-class assignments drawn from the book; you should bring it to every class.

The blog Language Log at `http://languagelog.ldc.upenn.edu/nll/` is also a "text" of sorts for this course; you should regularly at least skim it, and read articles that catch your eye. I will sometimes assign specific posts to read (mostly from LL but also occasionally other sources). Do check out the comments, which on this site do typically add value.

All programs in the course may be written in Python 3, using the IDLE development environment, and most in-class examples will be given using that language. If you are comfortable using other programming languages or environments, this will in general also be acceptable (check with me if you're not sure).

If you have an account on the math/CS department server you are welcome to use that for course assignments. If you don't have an account but are comfortable working in a command-line environment, I am happy to create one for you (but this is definitely not required).

Course objectives / Student learning outcomes

At the end of this course, the successful student will be able to:

1. write programs that analyse the content and structure of natural language text;
2. identify and explain linguistic properties of English and other world languages; and
3. draw data-driven quantitative conclusions about linguistic questions.

Core curriculum objective. In addition to the course-specific intended outcomes above, this course shares the objectives of the core curriculum as a whole. During this course, the successful student will:

4. develop and articulate informed perspectives essential to participation in civic and global life by integrating knowledge and skills across disciplines.

Core curriculum outcomes. Furthermore, during this course, the successful student will be able to:

5. locate, evaluate, and organize information from multiple disciplines to develop, refine, and address questions. — Information Literacy
6. use valid data and evidence from multiple disciplines to construct well-framed and well-supported arguments.
7. collaborate with others to develop an informed perspective on a civic or global issue.
8. reflect on the processes used to develop perspectives and reach decisions.
9. create and deliver writing appropriate to audience, purpose, and context. — Writing Infusion

Quantitative reasoning outcomes. As a Quantitative Reasoning course in the core curriculum, this course shares the following intended outcomes as well. At the end of this course, the successful student will be able to:

10. formulate a question/issue using appropriate mathematical, algorithmic, and/or statistical terms, and explain the decision process behind the choices made in that formulation;
11. use mathematical, algorithmic, and/or statistical methods to gather and/or analyze data—justification of the methods chosen should be included;
12. determine the reasonableness of an answer and/or evaluate the explanations of data for reasonableness, and understand the limitations behind the methods used in the previous outcome; and
13. interpret the results of a mathematical, algorithmic, and/or statistical analysis, and present the interpretation in a context appropriate for a broader audience.

Faculty objectives

Per section II-O-III-F-2 in the faculty manual, faculty teaching this writing-infused course are expected to:

14. integrate writing exercises and assignments with Core outcomes and individual course objectives, so that students may simultaneously master course content and develop their writing abilities;

15. provide explicit instruction to aid student understanding of writing appropriately for audiences in the relevant context or discipline; and

16. provide appropriate and timely peer and/or instructor feedback on student writing to allow opportunities for students to improve their writing through incorporating feedback on subsequent assignments.

Content

Graded work

I figure that I have on average about 9 hours of your time every week, including class time as well as reading, practice, homework, and projects. If you find you're regularly spending substantially more time than this, please do come discuss it with me, so that we can ensure you're making the most effective use of your time. The work you do for this course will be evaluated as follows:

Preparation and participation. You need to be an active participant in this class: present, prepared, and on-task. The grade for this component will be evaluated daily, using one of the following rubrics:

- Basic attendance: If you're there, you get the point!
- Participation: **1**: Attentive and on-task. **$\frac{1}{2}$ or 0**: Substantially late, sleeping, fussing with cellphone, etc.
- Reading quiz: Three questions, open-notes. **1**: Demonstrated that you read the assigned reading. **$\frac{1}{2}$**: Some correct work on the quiz.

You won't, in general, know in advance which I'll use on a particular day.

These points are worth 5% of the grade.

"Lab" programming. Many of the things we will discuss in this course will require some practice to try them out. I will from time to time assign shorter "lab" programming assignments, sometimes with an in-class programming component, due at the next class period. You're permitted to talk to your classmates about these assignments, although you'll have to submit your own work separately. Some will be purely

Collaborative CO 1

linguistic, some will be purely computer science, and many will be a mix of both. These will make up 10% of the grade.

Homework. From time to time, I will assign a short written homework at the end of class to be due at the beginning of the next. Each homework will proceed in two rounds: in response to your first handin, I'll give feedback (but no grade); after you have revised it, I'll assign a grade. Each problem will get 5, 3, or 0 points. The homeworks are group work: you can work with anyone in the class (or on your own if you prefer), and mark the names of the whole group at the top of a single handin. Some will be purely linguistic, some will be purely computer science, and many will be a mix of both. These will make up 10% of the grade.

Group
CO 2
CCO 7

Projects. A recurring aspect of the course will be that you will be called upon to formulate a question of interest and then do the work to answer it and write up your results. (See below for format and details.) These projects will form the central part of the course and thus a plurality of the course grade, 45%.

Collaborative
CO 1–3
CCO 4–9
QRO 10–13
Integrative

Exams. There will be two exams, one in late February and one during the finals period. The final will not be explicitly cumulative, though the material from the second half of the course will in some ways build on the earlier stuff. **You are not permitted to discuss the exams *at all,* with anyone other than me.** Each exam is worth 15% of the grade.

Non-collaborative
CO 1–3
CCO 4, 6

Grading scale

I tend to grade hard on individual assignments, but compensate for this in the final grades. The grading scale will be approximately as follows:

A−	[85, 90)	A	[90, 95)	A+	[95, 100]
B−	[70, 75)	B	[75, 80)	B+	[80, 85)
C−	[55, 60)	C	[60, 65)	C+	[65, 70)
D−	[40, 45)	D	[45, 50)	D+	[50, 55)

While there will be no "curve" in the statistical sense, I may slightly adjust the scale at the end of the term if it turns out some of the assignments were too difficult. Final grades of A+ are recorded as an A in the grading system. Final grades below the minimum for D− are recorded as an F.

Projects

The purpose of the course projects is to integrate the quantitative process with domain knowledge from the field of linguistics. In each, you will identify a linguistic question, explain why it is sociolinguistically relevant or interesting, and then try to answer it, using computational tools that you access and build: CO 1, 3 CCO 4–9

- First, you will formulate or reformulate the question into an explicitly quantifiable form; QRO 10
- then scrape or acquire the data from an appropriate source, and write and document a program to process the data and answer the quantified question; QRO 11
- and finally analyse and interpret that computational result to answer the original underlying question of interest. QRO 13
- Throughout the process, you will construct small, exact test cases to test your program, as well as using your understanding of English and other languages to evaluate the larger results for correctness and reasonableness. QRO 12

The questions (see below) are of a form often found in analysis and punditry, of politics or literature or just general social observation; but frequently the claims are made without support. What you will be doing is a form of citizen science: you'll be taking the claims and actually supporting them experimentally—or debunking them. CCO 4, 6

As part of that, you will be responsible for collecting your own data on the projects. We will have some discussions in class about what might serve as a reliable source for the linguistic data as well as how to go about acquiring the data itself in a way that will let it be processed computationally. CCO 5

During the course of the projects, you're encouraged to discuss your strategies and preliminary (and final!) results with other students in the course—they are what I refer to as "collaborative" assignments in my collaboration policy—which may help you refine your linguistic and computational insights as well as serving as an additional check on reasonableness. Other students will in general come from different disciplinary backgrounds as well as different language backgrounds; make use of that! CCO 7 QRO 12

At the end of the project, you will submit the program itself and your data (or a link to it) but the primary artifact you will produce is a written report on CCO 8, 9 FO 14

your work. As with many reports of this type, it has a dual audience: for those readers competent to do so, it should explain the full quantitative process in a way that would make the experiment reproducible; but it should also explain the conclusion in a way that would be understandable and persuasive to an interested audience that does not necessarily know or understand the intervening quantitative details.

Specifics

Project 1: Stake a claim of the general form "X talks a lot/not much about Y." X could be a person (such as Donald Trump or JK Rowling) or a group (such as Democrats or the Roman Catholic Church). Y is a topic. How will you defend and prove such a claim empirically and quantitatively?

Project 2: Answer a question of the general form "How real is prescriptive grammar rule X?" X should be an oft-cited rule (such as: no singular they, no split infinitive, which/that, no sentence-ending preposition, anything from Strunk and White—but in any case, give a citation for the rule!). You will use data from real and acknowledged-good writers (such as Jane Austen, Shakespeare, Dickens, Hemingway, or JK Rowling) to defend or debunk the rule. What will you count, and how will you make your argument?

Project 3: Explore data and draw a conclusion of the general form "Language learners from L1 X to L2 Y are particularly prone to make error Z." X and Y are languages, and for practical reasons one of them is probably English. Z can be any linguistic error. What sorts of statistics would be used to draw that kind of conclusion? What sorts of data would you need to gather to support it—how would you find it, clean it, and determine its usability?

The writing infusion

The course projects, each of which integrates a start-to-finish quantitative reasoning process with knowledge and reasoning in the linguistic domain, also serve as a recurring chance to develop your writing skills. FO 14

Many of the readings for this course will be online examples of just the sort of experimental linguistics you will be doing in your projects. As we lead into the first of the projects, I will explicitly lay out how to use these readings as a model for what you'll be writing in your own projects. FO 15

Following each of the first two projects, I will provide you with feedback not FO 16

only on the technical aspects of your project (the program itself and your analysis of the results) but also on the writing in your project report. This will give you a chance to improve your craft in the later ones.

The projects are, as noted above, collectively worth 45% of the grade.

Information literacy and critical thinking

> The trouble with this idea, as often with the insights of the punditocracy, is that there's no evidence that it's true. Worse, evidence is easily available to disconfirm it. —Mark Liberman[1]

While the field of computational linguistics is a broad one, and we will often hint at directions that the continuing student could take, much of the thematic inspiration for this course—and especially for the style and content of its projects—comes from the work of Mark Liberman and his collaborators at *Language Log*. A recurring theme on that blog is the idea that an educated citizen, armed with critical thinking, a little bit of programming background, and a little bit of linguistic sophistication, can often directly evaluate the plausibility or truth of "common sense" claims that are used to support supposedly insightful commentary.

Is an insight valid if it's founded on a fact that turns out to be false?

Critical thinking lets you read something and think, "wait, is that really true?" Information literacy, then, is the set of skills that lets you assemble and evaluate information to answer that question. Sometimes the best or only way to do so is to to consult a trusted authority; but an empowering alternative is to evaluate the truth of such a statement more directly, and it is this avenue that will be our focus in this course.

Calendar

Week 1

Overview of course
Review of programming basics
Finding good programming help on Stack Overflow and similar sites CCO 5

[1] Liberman, Mark. (2009, June 7). Fact-checking George Will [Blog post]. Retrieved from http://languagelog.ldc.upenn.edu/nll/?p=1486

Week 2

Quantitative reasoning process and the science of linguistics
Project 1 out Friday

Week 3

Ch1 Representing text and speech

Week 4

Ch2 Errors in speech and writing

Week 5

Ch2 Canonical forms, word lemmas
Soundex

Week 6

Ch2 Dynamic programming and edit distance
Spelling correction
Project 1 due Tuesday

Week 7

Ch2 Grammar, take 1
Exam Thursday **Exam 1, 28 February**
Project 2 out Thursday

SPRING BREAK

Week 8

Ch4 Text search and HTML

Week 9

Ch4 Regular expressions and pattern matching

Week 10

Ch4 Grammar, take 2
Ch7 Translation

Week 11

Ch7 Requirements and difficulties of translation
Project 2 due Tuesday
Project 3 out Thursday

Week 12

Ch7 IBM Model 1 alignment and translation

Week 13

Ch6 Dialogue and conversation

Week 14

(No class Tuesday)
Ch6 Dialogue systems

Week 15

(Tuesday only)
Ch6 Eliza
Proj 3 due Tuesday

The final exam is on Thursday, 2 May at 11:30am

Policies

You can find several university-wide course policies at
`http://www.longwood.edu/academicaffairs/syllabus-statements/` .

Support

This is an introductory course. That means that what is covered is an important basis for other work in the field, *not* that it is supposed to be obvious, or easy. So don't feel bad if something doesn't click right away. Never hesitate to ask me a question; I'll usually at least give you a hint as to where to look next.

I'm in my office a lot (not just during posted office hours). Feel free to come in and ask questions (or just to talk). If you can't catch me in my office, email is probably your best bet.

You should also make use of your fellow students as resources. While you can't copy each other's work (see the collaboration policy), studying together is a great idea, and asking and answering questions of other students is actively encouraged.

Accommodations

If you have any special need that I can accommodate, I'm happy to do so; come speak to me early in the term so we can set things up. If you have a documented disability, you should also contact Longwood's Office of Disability Resources (Brock Hall, x2391) to discuss some of the support the college can offer you. All such conversations are confidential.

Honor code policy

Above all, I ask and expect that you will conduct yourself with honesty and integrity—and not to ignore the other ten points of the Honor Code, either. Take pride in what you are capable of, and have the humility to give credit where it is due.

The two main forms of academic dishonesty are "cheating" and "plagiarism". "Cheating" is getting help from someplace you shouldn't, and "plagiarism" is presenting someone else's idea as if it's your own. If you ever find yourself

inclined towards either of these, know that there are always other, better options. Persevere! See my website[2] for some discussion and examples of how to steer clear of these problems, and feel free to come talk to me if you need help finding some of those other options (even if it's for another course).

Cheating or plagiarism (on any assignment) will normally receive a *minimum* penalty of a lowered *course* grade, ranging up to an F in the course. Cases will also be turned in to the Honor Board. But: I believe in your potential, and I hope that you will, or will grow to, observe this policy not simply to evade punishment but positively as a matter of character.

Attendance and late policy

Attendance is required, and assignments must be turned in on time. That said, if you have a good reason to miss class or hand something in late, I tend to be fairly liberal with extensions if you ask in advance. (Good reasons do include assignments due for other classes.) (And medical and family emergencies are exempted from the "in advance" part, of course. But contact me ASAP.)

Frequent absence will result in a lowered participation grade; habitual absence may in extreme cases result in a failing grade for the class. *Unexcused* late assignments will normally be given a zero.

Inclement weather policy

I don't plan to cancel class for weather unless the entire college shuts down. If you are commuting or are otherwise significantly affected by a weather event, use your own best judgement; and if you do miss class for this reason, contact me as soon as possible to make up missed work.

Early bird policy

Nobody's perfect, and on occasion an assignment gets written a little unclearly (or, once in a while, with an actual error in it). If you catch one and bring it to my attention early, so that I can issue a clarification or correction to the rest of the class, there'll be some extra credit in it for you.

[2] `http://cs.longwood.edu/~dblaheta/collab.html`