# Mathematical Methods in Linguistics

Fall 2023

| Course | Info |
|---|---|
| Course# | LIN 361/539 |
| Time | Tue/Thu 10:00-11:20am |
| Location | SBS S-228 |
| Website | lin361.thomasgraf.net |
| | lin539.thomasgraf.net |
| Instructor | Thomas Graf |
| Email | [coursenumber]@thomasgraf.net |
| Office hours | Tue 11:30am-12:30pm |
| | Wed 10:00am-11:00am |
| | Thu 11:30am-12:30pm |
| Office | SBS N-249 |
| TA | Han Li |
| TA email | han.li.4@stonybrook.edu |
| TA office hours | Mon 10:00-11:00am |
| | Wed 10:00-11:00am |
| TA office | SBS-N230 |

## 1 Course Outline

### 1.1 Bulletin Description

An overview of the mathematical foundations of theoretical and computational linguistics. Topics covered include set theory, morphisms, logic and model theory, algebra, lattices, lambda calculus, probability theory, information theory, and basics of formal language theory. A strong emphasis is put on the linguistic application of the mathematical concepts in the study and analysis of natural language data.

## 1.2 Full Description

This course is an introduction to mathematics in linguistics. It aims to help students familiarize themselves with mathematical concepts and applications that are widely relevant to theoretical and/or computational linguistics. This covers a wide range of topics, mostly from *discrete mathematics*. All the concepts will be illustrated with examples that have to do with natural language. For linguistics students, this should make the math more approachable and engaging. For mathematics students, this will show you a little-known but fascinating application area for mathematics.

The course is also very different from what you did in high school, there's precious few numbers here and we don't care much about trigonometry or calculating compound interest. In contrast to a proper mathematics course, we also focus more on techniques and tools rather than theorems and proofs. This means that you will learn how to work with things like functions, matrices, lattices, and finite-state automata, but you won't have to prove things about them. So this is more like a CS methods course than a proper math class.

For more information about the content, see the Selected Topics section. You will see that the schedule for this class is very ambitious. We probably won't be able to cover everything (I have taught this course many times, and I never made it all the way to the end). But no matter how far we get, by the end of the class you should have had enough exposure to mathematical thinking that the thought of picking up a textbook in mathematics or computational linguistics for self-study won't make you run away in horror.

## 1.3 Teaching Goals

- master essential concepts and techniques in mathematics and theoretical computer science
- apply mathematical techniques to the study of language
- formalize linguistic ideas in mathematical terms
- develop learning autonomy and the ability to expand your mathematical knowledge through self-study

## 1.4 Prerequisites

No prior mathematical or computational experience is required. We will get to see quite a bit of mathematical notation, like $(a \to \neg a) \to \neg a$ or

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \otimes \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}$$

But you don't need to know yet what any of that means.

There's no linguistic prereqs either, but since I'm a linguist I might sometimes forget what parts of linguistics are not common knowledge. Don't feel shy to ask clarification questions. Anecdotal evidence: I have run this course as an independent study with high school students who had no prior experience in linguistics, and they could follow along just fine because they were not afraid to ask me to clarify some linguistic concepts.

### 1.5 Textbook

None, but there are detailed lecture notes that will eventually become a textbook. The lecture notes will be made available as PDFs and HTML. Detailed instructions will follow in a separate announcement.

## 2 Selected Topics

A brief selection of the topics to be covered (we will probably deviate from this order):

1. Basic mathematical objects
    - Topics: sets, multisets, tuples
    - Applications: bag of words model of text, n-gram models of grammaticality
2. Grammatical systems as mappings
    - Topics: functions, monotonicity
    - Applications: monotonicity as a linguistic universal in morphology, syntax, and semantics
3. Relations and orders
    - Topics: properties of orders, posets, lattices, perhaps antimatroids
    - Applications: string extension learners, feature systems, adjunct algebras, syntactic relations, linguistic universals, OT if we cover antimatroids
4. Graph theory (we might skip this one)
    - Topics: (un)directed graphs, connectedness, components
    - Application: parse forest representation, autosegmental phonology, AVMs, unification grammars
5. Automata theory
    - Topics: finite-state automata and transducers
    - Application: upper bounds on the complexity of phonology & morphology
6. Linear algebra
    - Topics: vectors and vector spaces, matrices, tensor product
    - Application: automata as boolean matrix multiplication, vector space semantics, neural networks
7. Logic
    - Topics: propositional logic and first-order logic, types, lambda calculus
    - Application: semantics, model-theoretic syntax, subregular linguistics, CCG
8. Abstract algebra (only got to it once)
    - Topics: monoids, groups, semirings
    - Application: violation semirings in OT, semiring parsing
9. Types of infinity (usually not enough time)
    - Topics: bijections, function inverse
    - Applications: is language infinite?
10. Probability theory (extremely unlikely)
    - Topics: calculating probabilities with addition and multiplication
    - Application: weighted context-free grammars, corpus-based techniques
11. Information theory (extremely unlikely)

- Topics: entropy, cross-entropy
- Application: probabilistic machine learning, surprisal for processing

# 3 Grading

## 3.1 Undergraduate students

By default, 100% of your grade is determined by a (take-home) final exam. But there are various things you can do throughout the semester to reduce the importance of the final exam. This way, students who would rather study on their own can just do that and get their grade by demonstrating their knowledge on the final exam. Students who participate more regularly instead get to minimize the risk that a bad final exam will ruin their grade.

1. **Class participation (10%)**
   Show up all the time, ask questions and participate in discussion, and you will get an A for class participation. That A makes up 10% of your grade.

2. **Pre-assessment (P/F; 5%)**
   At the beginning of the semester, students are asked to take a survey to assess their prior knowledge of mathematical linguistics. It is perfectly normal not to know a single answer. Bring the completed survey to the first session of week 2. Performance is P/F depending on whether a filled-out survey was submitted (answering "Don't know" on each question is perfectly fine). If you get a P, that is treated like an A and is worth 5% of your grade.

3. **Weekly assignments (P/F; 3% each)**
   A list of exercises from the lecture notes is assigned every Thursday and your answers are due the following Tuesday. You should make a reasonable effort to complete the exercises, but your answers are not really graded. As long as it is clear that you made an effort, you get a P, which is treated as an A and is worth 3% of your grade. There will be around 10 assignments in total, making up 0% to 30% of your grade.

   Full solutions will be distributed each Thursday together with the next assignment.

4. **Take-home midterm (percentage graded; 30%)**
   We will have a midterm, but they won't be taken in class. Instead, the midterm takes the form of an extra-long homework assignment. If your grade on the midterm is higher than the grade on the final exam, it makes up 30% of your grade. If it is lower, we simply drop it and it does not factor into your grade (if you bomb the midterm but then shape up and turn in a great final exam, that's something to be rewarded, not punished).

5. **Post-assessment (P/F; 5%)**
   Exactly the same as the pre-assessment. Just do it again, pat yourself on the shoulder for every question you now know the answer to, and then hand it in with the final exam.

6. **Take-home final (percentage graded; remainder)**
   Same deal as the midterm, but even longer and you get two weeks to work on it.

## 3.2 Graduate students

The grading is the same as for undergrads, except that:

1. The weekly assignments may contain exercises that are optional for undergrads but mandatory for grads.
2. Once during the semester, you have to serve as pseudo-grader. You will get to look at half of the assignments, then you will meet with the TA to discuss what you think the students did well, what they struggled with, and who (if anybody) you think should not get a P due to lack of effort.

## 3.3 Student pseudonyms

Because the grad students in this course will get to look at the handed-in assignments, we will put in place an extra measure to protect your privacy. At the beginning of the semester, you will pick a pseudonym, and whenever you hand in an assignment, you will use that pseudonym instead of your real name. Detailed instructions will be sent out in a separate announcement.

## 3.4 Chat GPT

You are allowed to use Chat GPT for assignments and the final exam, but if you do so, please indicate that you did. It won't matter for grading or anything else, but I would like to know whether students actually benefit from using Chat GPT. Here is some advice based on my own experimentation:

1. Do not blindly copy-paste answers from Chat GPT. Treat it like a discussion with a peer, be aware that said peer may well be wrong, then write up your own answer.
2. Chat GPT will often get things wrong, in particular on the exercises that combine math and linguistics. But it can be insightful to think about what it got wrong, and why. Contrast what Chat GPT does against the official solutions for exercises, and you may have quite a few epiphanies.
3. Use Chat GPT selectively. It can take a lot of time to convert an exercise into a prompt that Chat GPT can do something with. It's often faster to just do things by yourself. Use Chat GPT for the exercises that really have you stymied, the exercises where you don't even know how to get started.

## 3.5 A remark on grades

Yes, this class has a grading system that is easy to game. You could do nothing the whole semester and then spend a day with Chat GPT to write up your take-home final. But that means you'll have paid the university a decent sum of money on a course that you learnt nothing from. If you're okay with that, then I'm okay with that, just like your gym doesn't mind that you pay them fifty bucks a month but never show up.

My experience is that the students who shouldn't get a good grade somehow manage to get a low grade even under the most lenient grading system. No reason to tighten clamps that will only make things more annoying for students who do just fine without all those extra hoops to jump through.

# 4 Policies

## 4.1 Contacting me

- Emails should be sent to [coursenumber]@thomasgraf.net. Disregarding this policy means late replies and might easily make me cross.
- Reply time < 24h in simple cases, possibly more if meddling with bureaucracy is involved.
- If you want to come to my office hours and anticipate a longer meeting, please email me so that we can set apart enough time and avoid collisions with other students.

## 4.2 Student Accessibility Support Center Statement

If you have a physical, psychological, medical, or learning disability that may impact your course work, please contact the Student Accessibility Support Center, Stony Brook Union Suite 107, (631) 632-6748, or at sasc@stonybrook.edu. They will determine with you what accommodations are necessary and appropriate. All information and documentation is confidential.

Students who require assistance during emergency evacuation are encouraged to discuss their needs with their professors and the Student Accessibility Support Center. For procedures and information go to the following website: https://ehs.stonybrook.edu//programs/fire-safety/emergency-evacuation/evacuation-guide-disabilities and search Fire Safety and Evacuation and Disabilities.

## 4.3 Academic Integrity Statement

Each student must pursue his or her academic goals honestly and be personally accountable for all submitted work. Representing another person's work as your own is always wrong. Faculty is required to report any suspected instances of academic dishonesty to the Academic Judiciary. Faculty in the Health Sciences Center (School of Health Technology & Management, Nursing, Social Welfare, Dental Medicine) and School of Medicine are required to follow their school-specific procedures. For more comprehensive information on academic integrity, including categories of academic dishonesty please refer to the academic judiciary website at http://www.stonybrook.edu/commcms/academic_integrity/index.html

## 4.4 Critical Incident Management

Stony Brook University expects students to respect the rights, privileges, and property of other people. Faculty are required to report to the Office of Student Conduct and Community

Standards any disruptive behavior that interrupts their ability to teach, compromises the safety of the learning environment, or inhibits students' ability to learn. Until/unless the latest COVID guidance is explicitly amended by SBU, during Fall 2021 "disruptive behavior" will include refusal to wear a mask during classes.

For the latest COVID guidance, please refer to: https://www.stonybrook.edu/commcms/strongertogether/latest.php