Subword Modeling

11-324/11-824

For up-to-date information, see the course webpage at
`https://dmort27.github.io/subwordmodeling/`.

1 Instructors

Instructor Prof. David R. Mortensen

Office Hours Tue 13:00–14:00, Wed 10:00-11:00

Office GHC 5407

Email dmortens@cs.cmu.edu

TA Anjali Kantharuban

Office Hours Mon 16:00–17:00, Wed 16:00–17:00

Location GHC 5417

Email akanthar@andrew.cmu.edu

2 Schedule

Do not panic: students are expected to do **one** reading for each class period (not the whole set of readings).

Table 1:

LEC	TOPIC	READINGS	DUE
	Module 1: Morphemes		
1	Introduction; introduction to Project 1	Haspelmath and Sims (2010) Ch 1–2 Park et al. (2021); Church (2020),	
2	Signs, minimal signs, and compositionality		
3	Productivity	Haspelmath and Sims (2010); Park et al. (2021), Ch 6	
4	Inflection, derivation, and compounding	Haspelmath and Sims (2010), Ch. 5; Matthews (1974), Ch. 5; Chaudhary et al. (2020); Hofmann et al. (2021)	
5	Morphotactics, affix ordering, the mirror principle, and the relevance principle	Aksënova et al. (2016)	

Continued on next page

Table 1: (Continued)

6	Computational approaches to morphological segmentation/tokenization (unsupervised [BPE, sentencepiece, Morfessor, etc.], supervized)	Sennrich et al. (2016); Mielke et al. (2021); Creutz and Lagus (2005); Virpioja et al. (2013); Khandagale et al. (2022); Bostrom and Durrett (2020); Hofmann et al. (2022)	
7	Allomorphy	Matthews (1974), Ch 6; Haspelmath and Sims (2010), Ch. 10; Yıldız et al. (2019)	
8	Non-concatenative processes	Amrhein and Sennrich (2021); Klein and Tsarfaty (2020); Fullwood and O'Donnell (2013); Haley and Wilson (2021)	
	Module 2: Lexemes		
9	Lexemes and paradigms; introduction to Project 2	Matthews (1974), Ch. 2; Haspelmath and Sims (2010), Ch. 8	
10	Grammatical properties	Matthews (1974), Ch. 9; Sylak-Glassman (2016)	Project 1
11	Word-and-Paradigm morphology	Matthews (1974), Ch 10	
12	Rules of realization and rules of referral: a computational system for WP morphology		
13	Neural approaches to reinflection	Cotterell et al. (2016); Pimentel et al. (2021)	
14	Neural approaches to paradigm completion	Jin et al. (2020); Wiemerslage et al. (2022)	
	Module 3: Graphemes		
15	Descriptive phonetics	International Phonetic Association (1999, 1–37)	
16	IPA versus orthographies. Introduction to Project 3		
17	Typology—alphabets, abjads, abugidas, syllabaries, logographic scripts, and others	Hockett et al. (1997)	Project 2
18	G2P and P2G	Mortensen et al. (2018); Li et al. (2022)	
19	Unicode—logical and visual representations	(Haralambous and Dürst, 2019)	

Continued on next page

Table 1: (Continued)

20	Input methods (ML and HCI perspectives)	van Esch et al. (2019)	
	Module 4: Sounds		
21	Articulatory features; introduction to Project 4	Mortensen et al. (2016), Li et al. (2021), Zouhar et al. (submitted)	Project 3
22	Syllabification and syllable segmentation	Bartlett et al. (2009); Mayer (2010)	
23	Phonological similarity and cognate detection	Bharadwaj et al. (2016); Chaudhary et al. (2018)	
24	Phonological representations and phonological alternations	Barke et al. (2019)	
25	Sound change	Ceolin and Sayeed (2019)	
26	Rule ordering and relative chronology	Boldsen and Paggio (2022)	Project 4
	Presentations		
27	Presentations		

3 Motivation

The goal of this course is to **lead students to engage broadly with the existing NLP and computational linguistics research on subword modeling** and **develop new computational approaches to problems in morphology, orthography, and phonology**. In addition to three other miniprojects, students will be expected to produce one piece of research that can be developed into a conference or workshop paper (though submission is not a course requirement). The paper should be suitable for the "Phonology, Morphology, and Word Segmentation" tracks of the *ACL conferences, the SIGMORPHON workshop, Coling, or LREC.

Natural Language Processing and Computational Linguistics have traditionally been biased towards phenomenon above the level of the word: syntax, semantics, and discourse. Contemporary neural models have been very successful in modeling many of these phenomena using words or information-theoretically defined subword units as atomic tokens. However, this has left the structure and patterns that exist inside of words—at the level of morphology, orthography, phonology, and phonetics—relatively unexplored. Of course, there has been computational work in all of these fields for decades but they have never been given the same degree of attention as syntax, semantics, and discourse.

Now, as NLP and CL become increasingly multilingual, interest in languages with richer phonologies, orthographies, and morphologies than English and Chinese has grown. A new wave of research, informed by both linguistic theory and machine learning methods, has opened up novel perspectives on problems in this domain and both practical and scientific issues that have not previously been explored. This course aims to place students on this threshold, reading to make new contributions in this burgeoning subfield.

4 Learning Objectives

At the end of this course, students will:

- Express a sophisticated understanding of the fundamentals of morphology, phonetics, writing systems, and phonology
- Understand current papers in computational phonology, morphology, and orthography
- Implement current subword models of language and carry out publication-quality experiments in this area
- Write publication-quality papers reporting computational research in phonology, morphology, and orthography

5 Mode of Instruction

Class meetings will be structured around handouts (not slides), which will also double as lecture notes (the instructors preferred mode of instruction). This allows a freer and more fluid discussion of the course content than a regimented slide presentation. Furthermore, handouts can provide a rich source of material for student's subsequent reference.

6 Assessments

6.1 Highlights (10 pts)

For 25 of 27 class meetings, each students will submit a list of three highlights: two takeaways from the day's lecture and one question. These highlights are intended to be brief (one sentence each) and to encourage engagement with the lecture. These must be submitted by 11:59pm the day of the corresponding class meeting.

6.2 Projects (60 pts)

The course will be assessed primarily through four mini-projects. Three of the four projects are structured as shared tasks. Students are allowed to work individually or in pairs. All teams are presented with:

- A dataset (`train`, `dev`, and `test`; `test` labels are held out)
- A metric, an evaluation script, and an autograder for evaluation via Gradescope
- A baseline (usually implementing some of the linguistic insight presented in the lectures in a rule-based model)

Students receive full credit if they beat the baseline on the `text` set. If they achieve the highest score on the task, they receive 5 bonus points. Unlimited submissions are allowed.
The G2P project is structured differently (see §6.2.3 below).
To receive credit, for each project a student must submit a two page report (ACL template) describing their experiements and the results (as well as referencing earlier work that has inspired theirs).

6.2.1 Morphological Segmentation (15 pts)

dataset Segmented data from two languages (Rarámuri and Shipibo-Konibo).

task Given training and development sets, tag previously unseen words with morpheme boundaries.

metric F1

baselines WFST-based (rule based) segmenter, Unigram tokenization (students must beat both for all languages)

6.2.2 Reinflection and Paradigm Completion (15 pts)

Task 1

dataset Three typologically diverse languages from Unimorph

task Given training and development sets, a previously unseen lemma, and a set of inflectional properties, return the inflected form of the word

metric Exact match accuracy, character error rate

baseline Rule-based word-and-paradigm inflector

Task 2

dataset Three typologically diverse languages from Unimorph

task Given training and development sets and a previously unseen partial paradigm, predict the other wordforms in that paradigm

metric Exact match accuracy, character error rate

baseline Rule-based paradigm completion engine ("rules of referral")

6.2.3 G2P for Three Languages (15 pts)

Students are required to implement rule-based or data-driven G2P for a language not supported by Epitran2 within the Epitran2 framework (including tests). Students can opt to have their modules included in subsequent uses of Epitran2 and to become coauthors on the Epitran2 paper (in progress).

6.2.4 Cognate Detection (15 pts)

Cognates are words in two or more languages that are descended from the same word in their shared ancestor language (like English *brother* and German *Bruder*).

dataset Mortensen-Wagel Comparative Tangkhulic Database

task In an unsupervised or zero-shot fashion, identify all cognacy relations between words in the four related languages

metric Precison@5 (reference is gold cognacy judgments)

baseline Cosine similarity between mean-pooled phonological feature vectors with empirically-determined threshold

6.3 Final Presentation and Paper (30 pts)

Students will select one of their projects (or a different idea of their choosing), develop it into a potentially publishable research product and make a conference-style oral presentation of it to the class. This presentation will be evaluated both on its technical and scientific merit and on its communicative effectiveness. The same research will be written up as four- to eight-page research paper which will be evaluated generously based on its likelihood to be accepted by one of the target venues.

7 Grading

Grades will be assigned as follows:

Assessment	Points
Highlights	10
Mini-Projects	60
Final Project	30

The grading scale will be as follows:

Grade	Range	
A+	100%	to 94.0%
A	< 94.0%	to 90.0%
A-	< 90.0%	to 87.0%
B+	< 87.0%	to 84.0%
B	< 84.0%	to 80.0%
B-	< 80.0%	to 77.0%
C+	< 77.0%	to 74.0%
C	< 74.0%	to 70.0%
C-	< 70.0%	to 67.0%
D+	< 67.0%	to 64.0%
D	< 64.0%	to 61.0%
F	< 61.0%	to 0.0%

Highlights will be based on reasonable completion. Mini-projects will be based on meeting the base line for each challenge. Scores below the baseline will credited according to a linear function. The student with the highest score on each tasks will receive 5 bonus points. The final paper will be based on one of the four mini-projects. The deliverable will be a paper and presentation, which will be graded together according to the following criteria:

1. Is related work handled appropriately? (10%)
2. Is the methodology well-motivated and technically sound? (20%)
3. Are appropriate baselines employed? (15%)
4. Are the experiments well designed? (15%)
5. Does the analysis of the results follow accepted statistical practices? (10%)
6. Is the paper well-written? (15%)
7. Was the structure and presentation of the talk clear and effective (15%)

8 Policies

8.1 Academic integrity

Any cheating or plagiarism will be dealt with according to the University policies on academic integrity. In general, discussion of tools, concepts, and formalisms is acceptable collaboration and is encouraged. Misrepresenting the work of others as your own, however, is considered cheating.

8.2 Late Policy

This course works best when everybody completes their work by the designated deadlines. This prevents cascading tardiness from overwhelming both students and teaching staff. However, sometimes there are situations that call for extensions. Some examples (real examples) include the following:

- The death of friend or family member
- A wedding in the family
- A serious accident
- A surgery
- A significant illness
- A mental health crisis or episode
- An important religious or national holiday

We care about you and your well being more than we care about deadlines and if something difficult is happening in your life which is making it hard for you to complete an assignment on time **please contact the instructor so you can talk**. We have found that, often, the students who most need some leeway are those least likely to ask for it. It never hurts to ask. We will work out a plan so you can complete the requirements of the course with your physical and psychological health intact. **Do not feel ashamed to reach out to us.** We are eager to see you succeed.

8.3 Equity and Inclusion Policy

Throughout human history, some people have been denied the rights and opportunities available to others on the basis of their race, gender, economic class, caste, ancestry, language community, age, religion, beliefs, political affiliation, and abilities (visible and invisible). A single course cannot undo the injustices of history, but we—as a teaching staff—are committed to fighting inequity and promoting inclusion. We encourage you to join us. If you feel that you, or those around you, have been treated unfairly based upon their identity (or perceived identity) by us, by other members of the teaching staff, or by other students in the course, we ask that you bring it to our attention so that we can address the wrongs (as well as pursing the approved University channels).

8.4 Disability Rights

Many people have disabilities, including members of our own families. We see disabilities as deficits not in disabled people but in the institutions and societies that are structured such that they are disadvantaged. We wish to do our part to overcome this disparate treatment. If you have a disability (visible or invisible), please let us know as soon as possible (you don't need to tell us the nature of the disability) and work with Disability Service to develop a set of accommodations which we can then approve. These might, for example, include lecture materials that are usable by people with visual disabilities, sign language interpretation, captioning, flexible due dates, etc.

References

Alëna Aksënova, Thomas Graf, and Sedigheh Moradi. 2016. Morphotactics as tier-based strictly local dependencies. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 121–130, Berlin, Germany. Association for Computational Linguistics.

Chantal Amrhein and Rico Sennrich. 2021. How suitable are subword segmentation strategies for translating non-concatenative morphology? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 689–705, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shraddha Barke, Rose Kunkel, Nadia Polikarpova, Eric Meinhardt, Eric Bakovic, and Leon Bergen. 2019. Constraint-based learning of phonological processes. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6176–6186, Hong Kong, China. Association for Computational Linguistics.

Susan Bartlett, Grzegorz Kondrak, and Colin Cherry. 2009. On the syllabification of phonemes. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 308–316, Boulder, Colorado. Association for Computational Linguistics.

Akash Bharadwaj, David Mortensen, Chris Dyer, and Jaime Carbonell. 2016. Phonologically aware neural model for named entity recognition in low resource transfer settings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1472, Austin, Texas. Association for Computational Linguistics.

Sidsel Boldsen and Patrizia Paggio. 2022. Letters from the past: Modeling historical sound change through diachronic character embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6713–6722, Dublin, Ireland. Association for Computational Linguistics.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

Andrea Ceolin and Ollie Sayeed. 2019. Modeling markedness with a split-and-merger model of sound change. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 67–70, Florence, Italy. Association for Computational Linguistics.

Aditi Chaudhary, Antonios Anastasopoulos, Adithya Pratapa, David R. Mortensen, Zaid Sheikh, Yulia Tsvetkov, and Graham Neubig. 2020. Automatic extraction of rules governing morphological agreement. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5212–5236, Online. Association for Computational Linguistics.

Aditi Chaudhary, Chunting Zhou, Lori Levin, Graham Neubig, David R. Mortensen, and Jaime Carbonell. 2018. Adapting word embeddings to new languages with morphological and phonological subword representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3285–3295, Brussels, Belgium. Association for Computational Linguistics.

Kenneth Ward Church. 2020. Emerging trends: Subwords, seriously? *Natural Language Engineering*, 26(3):375–382.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared Task—Morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2005. *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0*. Helsinki University of Technology Helsinki.

Daan van Esch, Elnaz Sarbar, Tamar Lucassen, Jeremy O'Brien, Theresa Breiner, Manasa Prasad, Evan Crew, Chieu Nguyen, and Franccoise Beaufays. 2019. Writing across the world's languages: Deep internationalization for gboard, the google keyboard. *ArXiv*, abs/1912.01218.

Michelle Fullwood and Tim O'Donnell. 2013. Learning non-concatenative morphology. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 21–27, Sofia, Bulgaria. Association for Computational Linguistics.

Coleman Haley and Colin Wilson. 2021. Deep neural networks easily learn unnatural infixation and reduplication patterns. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 427–433, Online. Association for Computational Linguistics.

Yannis Haralambous and Martin Dürst. 2019. Unicode from a linguistic point of view. In *Proceedings of Graphemics in the 21st Century, Brest 2018*, pages 167–183, Brest. Fluxus Editions.

Martin Haspelmath and Andrea Sims. 2010. *Understanding Morphology*, 2nd edition. Hodder Education, London.

Charles Francis Hockett, Peter T. Daniels, and William Bright. 1997. The world's writing systems. *Language*, 73:379.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves BERT's interpretation of complex words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3594–3608, Online. Association for Computational Linguistics.

Valentin Hofmann, Hinrich Schuetze, and Janet Pierrehumbert. 2022. An embarrassingly simple method to mitigate undesirable properties of pretrained language model tokenizers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–393, Dublin, Ireland. Association for Computational Linguistics.

International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

Huiming Jin, Liwei Cai, Yihui Peng, Chen Xia, Arya McCarthy, and Katharina Kann. 2020. Unsupervised morphological paradigm completion. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6696–6707, Online. Association for Computational Linguistics.

Katharina Kann, Jesus Manuel Mager Hois, Ivan Vladimir Meza-Ruiz, and Hinrich Schütze. 2018. Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 47–57, New Orleans, Louisiana. Association for Computational Linguistics.

Sujay Khandagale, Yoann Léveillé, Samuel Miller, Derek Pham, Ramy Eskander, Cass Lowry, Richard Compton, Judith Klavans, Maria Polinsky, and Smaranda Muresan. 2022. Towards unsupervised morphological analysis of polysynthetic languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 334–340, Online only. Association for Computational Linguistics.

Stav Klein and Reut Tsarfaty. 2020. Getting the ##life out of living: How adequate are word-pieces for modelling complex morphology? In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 204–209, Online. Association for Computational Linguistics.

Xinjian Li, Juncheng Li, Florian Metze, and Alan W. Black. 2021. Hierarchical Phone Recognition with Compositional Phonetics. In *Proc. Interspeech 2021*, pages 2461–2465.

Xinjian Li, Florian Metze, David Mortensen, Shinji Watanabe, and Alan Black. 2022. Zero-shot learning for grapheme to phoneme conversion with language ensemble. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2106–2115, Dublin, Ireland. Association for Computational Linguistics.

P. H. Matthews. 1974. *Morphology*. Cambridge University Press, Cambridge.

Thomas Mayer. 2010. Toward a totally unsupervised, language-independent method for the syllabification of written texts. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 63–71, Uppsala, Sweden. Association for Computational Linguistics.

Sabrina J. Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y. Lee, Benoît Sagot, and Samson Tan. 2021. Between words and characters: A brief history of open-vocabulary modeling and tokenization in nlp.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484, Osaka, Japan. The COLING 2016 Organizing Committee.

Hyunji Hayley Park, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021. SIGMORPHON 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, Mikko Kurimo, et al. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline.

Adam Wiemerslage, Miikka Silfverberg, Changbing Yang, Arya McCarthy, Garrett Nicolai, Eliana Colunga, and Katharina Kann. 2022. Morphological processing of low-resource languages: Where we are and what's next. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 988–1007, Dublin, Ireland. Association for Computational Linguistics.