Recent Developments in Computational Typology and Multilingual Natural Language Processing

June 16, 2021 · Issue #12

Editors: Ekaterina Vylomova, Pranav A, Eleanor Chodroff, and Ryan Cotterell

This is SIGTYP's twelfth newsletter on recent developments in computational typology and multilingual natural language processing. Each month, various members of SIGTYP will endeavour to summarize recent papers that focus on these topics. The papers or scholarly works that we review are selected to reflect a diverse set of research directions. They represent works that the editors found to be interesting and wanted to share. Given the fast-paced nature of research in our field, we find that brief summaries of interesting papers are a useful way to cut the wheat from the chaff.

We expressly encourage people working in computational typology and multilingual NLP to submit summaries of their own research, which we will collate, edit and announce on SIGTYP's website. In this issue, for example, we had Johann-Mattis List, Hiram L. Smith, Jakub Szymanik, Natalia Levshina, Ruochen Xu, Fajri Koto, Aryaman Arora, Anna Rogers, Khuyagbaatar Batsuren, Badr M. Abdullah, Tanel Alumäe describe their recent research on linguistic typology and multilingual NLP.

Research Papers **3**
- The Uses and Abuses of Tree Thinking in Cultural Evolution 3
- Towards a Sustainable Handling of Interlinear-glossed Text in Language Documentation 3
- Reflex Prediction. A Case Study of Western Kho-Bwa 4
- Do Creoles Conform to Typological Patterns? Habitual Marking in Palenquero 4
- Quantifiers Satisfying Semantic Universals are Simpler 5
- Why We Need a Gradient Approach to Word Order 6
- Mixed-Lingual Pre-training for Cross-lingual Summarization 7
- Evaluating the Efficacy of Summarization Evaluation across Languages 8
- Bhāṣācitra: Visualising the Dialect Geography of South Asia 8
- Changing the World by Changing the Data 9
- A Large and Evolving Cognate Database 10
- Rediscovering the Slavic Continuum in Representations Emerging from Neural Models of Spoken Language Identification 11
- SIGTYP 2021 Shared Task on Robust Spoken Language Identification: an Overview 11
- Do Acoustic Word Embeddings Capture Phonological Similarity? An Empirical Study 12
- Familiar Words but Strange Voices: Modeling the Influence of Speech Variability on Word Recognition 13

Books **13**
- Finite-State Text Processing 13

Resources **14**
- EDICTOR. A Web-based Interactive Tool for Creating and Editing Etymological Datasets 14
- VoxLingua107: A Speech Dataset for Training Spoken Language Identification Models 14
- Materials for Goals and Methods of Computational Linguistics 15

Discussions **15**
- What does “Native speaker” Mean, Anyway? (by Devin Grammon and Anna Babel) 15

Talks **15**
- SIGTYP Lectures 15
- SIGTYP on Youtube 15
- SIGTYP 2021 -- All Talks, Papers, and Discussions are Publicly Available! 16
- Abralin ao Vivo – Linguists Online 16

Research Papers

The Uses and Abuses of Tree Thinking in Cultural Evolution

Cara L. Evans, Simon J. Greenhill, Joseph Watts, Johann-Mattis List, Carlos A. Botero, Russell D. Gray and Kathryn R. Kirby

Summary by Johann-Mattis List

Modern phylogenetic methods are increasingly being used to address questions about macro-level patterns in cultural evolution. These methods can illuminate the unobservable histories of cultural traits and identify the evolutionary drivers of trait change over time, but their application is not without pitfalls. Here, we outline the current scope of research in cultural tree thinking, highlighting a toolkit of best practices to navigate and avoid the pitfalls and 'abuses' associated with their application. We emphasize two principles that support the appropriate application of phylogenetic methodologies in cross-cultural research: researchers should (1) draw on multiple lines of evidence when deciding if and which types of phylogenetic methods and models are suitable for their cross-cultural data, and (2) carefully consider how different cultural traits might have different evolutionary histories across space and time. When used appropriately phylogenetic methods can provide powerful insights into the processes of evolutionary change that have shaped the broad patterns of human history.

Towards a Sustainable Handling of Interlinear-glossed Text in Language Documentation

Johann-Mattis List, Nathaniel A. Sims, Robert Forkel

Summary by Johann-Mattis List

While the amount of digitally available data on the worlds' languages is steadily increasing, with more and more languages being documented, only a small proportion of the language resources produced are sustainable. Data reuse is often difficult due to idiosyncratic formats and a negligence of standards that could help to increase the comparability of linguistic data. The sustainability problem is nicely reflected in the current practice of handling interlinear-glossed text, one of the crucial resources produced in language documentation. Although large collections of glossed texts have been produced so far, the current practice of data handling makes data reuse difficult. In order to address this problem, we propose a first framework for the computer-assisted, sustainable handling of interlinear-glossed text resources. Building on recent standardization proposals for word lists and structural datasets, combined with state-of-the-art methods for automated sequence comparison in historical linguistics, we show how our workflow can be used to lift a collection of

interlinear-glossed Qiang texts (an endangered language spoken in Sichuan, China), and how the lifted data can assist linguists in their research.

Reflex Prediction. A Case Study of Western Kho-Bwa

Timotheus A. Bodt, Johann-Mattis List

Summary by Johann-Mattis List

While analysing lexical data of Western Kho-Bwa languages of the Sino-Tibetan or Trans-Himalayan family with the help of a computer-assisted approach for historical language comparison, we observed gaps in the data where one or more varieties lacked forms for certain concepts. We employed a new workflow, combining manual and automated steps, to predict the most likely phonetic realisations of the missing forms in our data, by making systematic use of the information on sound correspondences in words that were potentially cognate with the missing forms. This procedure yielded a list of hypothetical reflexes of previously identified cognate sets, which we first preregistered as an experiment on the prediction of unattested word forms and then compared with actual word forms elicited during secondary fieldwork. In this study we first describe the workflow which we used to predict hypothetical reflexes and the process of elicitation of actual word forms during fieldwork. We then present the results of our reflex prediction experiment. Based on this experiment, we identify four general benefits of reflex prediction in historical language comparison. These comprise (1) an increased transparency of linguistic research, (2) an increased efficiency of field and source work, (3) an educational aspect which offers teachers and learners a wide plethora of linguistic phenomena, including the regularity of sound change, and (4) the possibility of kindling speakers' interest in their own linguistic heritage.

Do Creoles Conform to Typological Patterns? Habitual Marking in Palenquero

Hiram L. Smith

Summary by Hiram L. Smith

The Creole Debate is, essentially, the question of whether or not creole languages constitute a typological class based on structural properties. However, generalizations from functional-typological literature have never been empirically tested. For example, typological studies report a strong cross-linguistic tendency for asymmetries in habitual expressions across present and past tense, and that habituals are "highly affected by tense", and that the default meanings of present and past tense are responsible for these formal asymmetries (e.g., whether a verb uses overt or zero coding to express habitual meaning) (Bybee 1994; Bybee et al.1994). On the other hand, traditional creole studies have reported that creole TMA systems, and in particular, their

preverbal markers, which have specific syntactic and functional roles that distinguish creole from noncreole varieties (Bickerton 1975, 1981, 1984). How does Palenquero, an Afro-Hispanic creole spoken in northern Colombia, fit into this debate?

Under the microscope are two linguistic variants, preverbal *asé* and zero, illustrated in example (1), which compete for habitual marking in Palenquero. I ask: To what degree does the patterning of these forms conform to well-attested cross-linguistic tendencies? How is habitual marking situated within the overall architecture of present temporal reference?

1) Ma jende **asé komblá** kottía y ma jende **Ø asé** un poko kumina.
PL people **HAB buy** ribs and PL people **make** a little food
'The people buy ribs and the people make a little food.' (Male 54, Recording 6, 5:32)

To answer these questions, quantitative analyses were conducted on 2,543 tokens of past and present tense forms in order to implement a methodology to test the applicability of crosslinguistic generalizations found in functional-typological literature to the Creole Debate. Results show that the asymmetrical expression, distribution, and relative ordering of forms closely align with typological predictions for habituals, thus giving convincing evidence of typological markedness and not a Creole Prototype.

An important contribution of this project was that it enabled systematic quantitative analysis of natural speech community data from an exhaustively transcribed corpus. I demonstrate that, while creoles are often typologized based on their structural properties, it is crucial to look, not only at structure, but the close relationship between form, distributions, and external function. This study hopes to marry such disparate areas as variationist, typological, and creole studies. It also aims to counter the stigmatization of Afro-Hispanic varieties by showing the systematicity of this creole language's preverbal particles, whose distributions adhere to cross-linguistic typological markedness patterns.

Quantifiers Satisfying Semantic Universals are Simpler

Iris van de Pol, Paul Lodder, Leendert van Maanen, Shane Steinert-Threlkeld, Jakub Szymanik

Summary by Jakub Szymanik

Despite wide variation among natural languages, there are linguistic properties thought to be universal to all or almost all natural languages. Maybe most famously such universals have been proposed by Barwise & Cooper in the semantic domain of quantification. They have observed that the quantifiers that are lexicalized (as monomorphemic words) in natural language share certain structural properties, namely those of monotonicity, quantity, and conservativity. However, when looking at the space of all logically possible quantifiers, a large majority does not have these properties. Hence, the question arises: why do these universals hold? Why do quantifiers in natural language have precisely these properties? A possible explanation for these universals lies in the

interaction between these properties and our cognitive apparatus. This builds on the idea that the human cognitive system is structured in such a way that it favors dealing with certain meanings over others. In this paper, we explore the following hypothesis: quantifiers having these semantic properties are simpler. Simplicity as an explanatory concept in cognition has been studied in a variety of domains. Here we measure simplicity by the concept of minimal description length in a logical grammar (i.e., a language of thought), a framework that has been used, e.g., in the domain of concept learning, language acquisition, and auditory memory. We use this grammar to generate a collection of over 24,000 logically possible quantifiers, and we compute both their complexities and their adherence to universal properties. Using a logistic regression model, we show that quantifiers with universal properties are simpler.

Why We Need a Gradient Approach to Word Order

Natalia Levshina, Savithry Namboodiripad, Marc Allassonnière-Tang, Mathew A. Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez Timothy Gupton, Evan Kidd, Zoey Liu, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova, Natalia Stoynova

Summary by Natalia Levshina

In this paper we argue for a gradient approach to word order. A gradient approach means that word order patterns should be treated as a continuous variable. For example, instead of labelling a language as SO (subject-object) or OS (object-subject), we can compute and report the proportion of SO and OS based on behavioural data (corpora and experiments). Similarly, instead of speaking of fixed, flexible, or free word order, we measure the degrees of this variability by using quantitative measures, such as entropy. In our paper we take a broad interdisciplinary perspective, considering diverse factors related to language acquisition, processing, language contact, language change and prosody, which lead to gradience and variability in word order. We also show how different language sciences can benefit from taking a gradient approach more seriously.

Despite some previous research problematizing categorical distinctions in linguistics (e.g. Wälchli 2009), non-gradient approaches to word order are still hegemonic. Yet, the situation is changing. There is a growing interest in probabilistic gradient phenomena which are captured by complex statistical models of language users' behaviour. In this paper we argue that non-gradient approaches are theoretically problematic and lead to loss of data.

Among many other things, we address several methodological challenges one faces when using corpus data to study word order from a gradient perspective:

1. Corpus size. How much corpus data are needed to obtain reliable results? We discuss a case study of S-V and O-V orders in several UD corpora, which demonstrates that most languages reach a relative stability of entropy after a sample size of 500 sentences.

2. Parsing errors. How dangerous are parsing errors? A case study of Shannon entropy of the relative order of nominal heads and modifiers in an automatically annotated parallel corpus of Indo-European prose and the corresponding UD treebanks reveals that the differences between the entropy scores in the datasets are not statistically significant.

3. Register and modality biases. How much do register and modality influence our results? A comparison of dependency-parsed YouTube captions in seven languages and the corresponding written UD corpora shows that the difference in dependency length minimization across speech and writing varies in a consistent way. This means that we should avoid generalizing to registers and modalities not represented in our corpora.

4. Phylogenetic methods. We also show that word order entropy estimates derived from the UD corpora in 36 Indo-European languages can be used for capturing phylogenetic signal with the help of phylogenetic methods.

Mixed-Lingual Pre-training for Cross-lingual Summarization

Ruochen Xu, Chenguang Zhu, Yu Shi, Michael Zeng, Xuedong Huang

Summary by Ruochen Xu

Cross-lingual Summarization (CLS) aims at producing a summary in the target language for an article in the source language. Traditional solutions employ a two-step approach: either translate the article into the target language and then summarize it, or summarize the article in the source language and then translate it. Although this method can leverage off-the-shelf summarization and MT models, it suffers from error accumulation from two independent subtasks. Recently, end-to-end models have achieved better results, but these approaches are mostly limited by their dependence on large-scale labeled data. Built upon a transformer-based encoder-decoder architecture, our model is pre-trained on both monolingual tasks including masked language model (MLM), denoising autoencoder (DAE) and monolingual summarization (MS), and cross-lingual tasks such as cross-lingual masked language model (CMLM) and machine translation (MT). This mixed-lingual pre-training scheme can take advantage of massive unlabeled monolingual data to improve the model's language modeling capability, and leverage cross-lingual tasks to improve the model's cross-lingual representation. Moreover, the architecture has no task-specific components, which saves memory and increases optimization efficiency. We show in experiments that this pre-training scheme can effectively boost the performance of cross-lingual summarization. In Neural Cross-Lingual Summarization (NCLS) dataset, our model achieves an improvement of 2.82 (English

to Chinese) and 1.15 (Chinese to English) ROUGE-1 scores over state-of-the-art results. The performance gain from pre-training is even larger when the finetune data is extremely small. We further conduct an ablation study to show that each pretraining task contributes to the performance, especially our proposed unsupervised pre-training tasks.

Evaluating the Efficacy of Summarization Evaluation across Languages

Fajri Koto, Jey Han Lau, Timothy Baldwin

Summary by Fajri Koto

While automatic summarization evaluation methods developed for English are routinely applied to other languages, this is the first attempt to systematically quantify their panlinguistic efficacy. We take a summarization corpus for eight different languages, namely English (EN), Indonesian (ID), French (FR), Turkish (TR), Mandarin Chinese (ZH), Russian (RU), German (DE), and Spanish (ES), and manually annotate generated summaries for focus (precision) and coverage (recall). We use two modern summarization systems for each language: pointer generator and BERT-based models, which result in 4,320 annotations. Based on this, we evaluate an extensive range of traditional and model-based metrics (19 in total), and find BERTScore (with mBERT uncased) to be the best metric for evaluating both focus and coverage universally.

Bhāṣācitra: Visualising the Dialect Geography of South Asia

Aryaman Arora, Adam Farris, Gopalakrishnan R, Samopriya Basu

Summary by Aryaman Arora

Bhāṣācitra (https://aryamanarora.github.io/bhasacitra/) is an online dialect mapping system for South Asia built on a database of linguistic studies of languages of the region manually annotated for topic and location data. We collected the database by scouring existing bibliographies, online scholarly repositories, and digitised archival materials. The application is not only meant to be useful for feature mapping, but also serves as a new kind of interactive bibliography for linguists of South Asian languages.

The main contribution of Bhāṣācitra is a map aggregating the location information we collected to delineate the geographical reaches of individual lects. We were inspired by the principles employed in other approaches to linguistic data visualisation, such as ParHistVis (Kalouli et al., 2019). Existing bibliographic maps like Glottolog (Hammarström et al., 2021) do not present areal information for

language distributions. In South Asia especially, the geographical reach of language varieties is contested and subject to sociolinguistic variation (e.g. multiple names used for the same language by different groups), which complicates official censuses and localised language surveys (Asher, 2008). By aggregating the linguistic literature at a fine level of geographical detail we establish a more complete picture of the language situation in the region.

Using the map, we look towards applications to typology by visualising example datasets. The vast linguistic diversity of South Asia results in big datasets; while statistical techniques have matured for studying typology, human interpretation of language data is still difficult. To demonstrate the value of geographical visualisation, we mapped the presence of /ɖʱ/ in phoneme inventories of 62 major South Asian languages (Ramaswami, 1999) and the probabilities of outcomes of sound changes of the Sanskrit cluster /kṣ/ in modern Indo-Aryan languages (Turner, 1962–1966). In both cases, we found apparent patterns that represent avenues for future typological study: the lack of /ɖʱ/ in languages of the southern, northwestern, and eastern fringes of the Subcontinent, and a preference for /kṣ/ > /kʰ/ in Northwest Indo-Aryan with gradual reduction of that preference towards the southern and eastern regions.

We hope to see more work in the creation of tools and the development of datasets for South Asia that benefit all linguists, as we study its vast linguistic diversity and develop language technologies for the region's languages and their speakers. To that end, we welcome contributions to our open-source repository at https://github.com/aryamanarora/bhasacitra. Ultimately, this approach to language mapping would be useful globally, so we would be happy to work with collaborators with domain knowledge of languages outside of South Asia.

Changing the World by Changing the Data

Anna Rogers

Summary by Anna Rogers

One of the biggest debates sparked by "On the Dangers of Stochastic Parrots" centered on data curation: do we want our models to represent the world "as it is", or do we want to try to improve the world by curating our data? A new position paper (to appear at ACL 2021) maps out the debate and argues that the point is moot: data curation is and will be happening, and it will change the world. The only question is whether we want to at least try to steer that process.

The key argument is that the only way to not change the world is if our models represent the world "as it is". But that is not the case: our data has both social and linguistic biases, which are *different* from the real-world biases. This makes any sampling procedure (even if we use all of Common Crawl) an implicit curation decision. And that decision has real consequences, because the conceptual and linguistic repertoire of human speakers is dynamic. Our language and ideas change

throughout our lifetime, influenced by ideas and language of other humans that we interact with -- especially those that we encounter more often.

Consider that from now on much of the speech we encounter in daily life will be synthetic -- simply because NLP models can be commercially deployed on a large scale. With all the schools and theaters in the world, Shakespeare probably has not had as much audience in the past 4 centuries as OpenAI's GPT3 has already had since last year: it currently generates 4.5 billion words a day in customer support, games, question answering and other applications. The linguistic and social patterns in this incredible volume of synthetic speech are going to impact both how we speak and how we think. They are also going to impact our actual lives: e.g. if a model assisting with student essay grading or job applications "understands" certain syntactic patterns better than others, or encodes implicit social biases, it can directly harm individuals or even social groups.

The bottom line is that the world is already changing because of data decisions we make, explicitly or implicitly. So the question is only whether we want to at least try to steer that process.

A Large and Evolving Cognate Database

Khuyagbaatar Batsuren, Gábor Bella, Fausto Giunchiglia

Summary by Khuyagbaatar Batsuren

We present **CogNet**, a large-scale database of cognate pairs: it contains 8.1 million cognates in 338 languages, 38 writing systems, and 91285 concepts. It was automatically constructed from wordnets and dictionaries contained within the UKC resource, as described in our paper.

What are cognates? In short, cognates are words in different languages that share a common origin and the same meaning, such as the English letter and the French letter. CogNet links its cognates to Princeton WordNet synsets, making the shared meanings explicit. Inside CogNet, two well-distinguished kinds of cognates are included:

- words that have the same meaning and are etymologically related according to gold-standard evidence (such as the Etymological WordNet or Wiktionary): such cognates, which constitute about 40% of CogNet, can be used for applications in historical linguistics;
- words that have the same meaning but we only have indirect evidence of their relatedness, such as orthographic or phonetic similarity. This corresponds to a more relaxed interpretation of cognacy that also includes, e.g., loanwords, and that is well suited for applications in computational linguistics such as machine translation or bilingual lexicon induction.

Why are cognates important? Cognates have been extensively studied in the fields of language typology and historical linguistics, as they are considered useful for researching the relatedness of

languages. This paper conducts a quantitative analysis of cognate diversity of concepts, leading to novel insights on language diversity.

Rediscovering the Slavic Continuum in Representations Emerging from Neural Models of Spoken Language Identification

Badr M. Abdullah, Jacek Kudera, Tania Avgustinova, Bernd Möbius, Dietrich Klakow

Summary by Badr M. Abdullah

Deep neural networks have been successfully employed for various spoken language recognition tasks, including tasks that are multilingual by definition such as spoken language identification (SLID). Most recent state-of-the-art SLID systems are trained end-to-end to map spectral representations of (untranscribed) speech onto high-level feature representations in a vector space where languages are linearly separable. These systems have shown impressive performance not only in discriminating between distant languages but also between closely related language varieties. However, it remains unknown whether the emergent distance in the representation space correlates with objective measures of language similarity. In our work, we present a case study on Slavic language identification and conduct a linguistically motivated analysis on the emergent language representations. Concretely, we make two key contributions: (1) We present an SLID model for Slavic languages with domain-adversarial training to improve the model robustness against non-language sources of variability in speech. (2) We analyze and visualize the emergent representations from our robust SLID model for 11 Slavic languages, five of which are not observed (held-out) during training. We show that the distance in the representation space correlates with both geographic and phylogenetic distances. Our work sheds light on the speech modality and presents a case study on how data-driven approaches that only require (untranscribed) speech segments can complement ongoing research efforts in the field of linguistic typology and language variation.

SIGTYP 2021 Shared Task on Robust Spoken Language Identification: an Overview

Elizabeth Salesky, Badr M. Abdullah, Sabrina Mielke, Elena Klyachko, Oleg Serikov, Edoardo Maria Ponti, Ritesh Kumar, Ryan Cotterell, Ekaterina Vylomova

While language identification is a fundamental speech and language processing task, for many languages and language families it remains a challenging task. For many low-resource and endangered languages this is in part due to resource availability: where larger datasets exist, they

may be single-speaker or have different domains than desired application scenarios, demanding a need for domain and speaker-invariant language identification systems. This year's shared task on robust spoken language identification sought to investigate just this scenario: systems were to be trained on largely single-speaker speech from one domain, but evaluated on data in other domains recorded from speakers under different recording circumstances, mimicking realistic low-resource scenarios. We see that domain and speaker mismatch proves very challenging for current methods which can perform above 95% accuracy in-domain, which domain adaptation can address to some degree, but that these conditions merit further investigation to make spoken language identification accessible in many scenarios.

Do Acoustic Word Embeddings Capture Phonological Similarity? An Empirical Study

Badr M. Abdullah, Marius Mosbach, Iuliia Zaitova, Bernd Möbius, Dietrich Klakow

Summary by Badr M. Abdullah

Several variants of deep neural networks have been successfully employed for building parametric models that project variable-duration spoken word segments onto fixed-size vector representations, or acoustic word embeddings (AWEs). In previous work, it has been hypothesized that the distance in the emergent AWE space can be interpreted as a metric of (perceptual) dissimilarity between linguistic units (e.g., phones, syllables, words). However, none of the previous studies has empirically (in)validated this hypothesis with a rigorous evaluation beyond the conventional intrinsic evaluations based on word discrimination. In this paper, we ask: does the distance in the acoustic embedding space correlate with phonological dissimilarity? To answer this question, we empirically investigate the performance of supervised approaches for AWEs with different neural architectures. Concretely, we conduct a set of experiments using word-aligned speech data from two languages (German and Czech) where we keep the training conditions for each model fixed and systematically study the impact of, and the interplay between, the architecture (convolutional and recurrent) and learning objective (classification, phonological decoding, and contrastive objectives) on model performance. We then analyze the correlation between the distance in the embedding space of each model and word-form (dis)similarity, which we measure using a phonetically-informed extension of Levenshtein distance. Our experiments show that (1) the distance in the embedding space in the best cases only moderately correlates with phonological distance, and (2) improving the performance on the conventional word discrimination task does not necessarily yield models that better reflect word phonological similarity. We also find that recurrent models which are trained with symbolic grounding are better at capturing word-form similarity compared to their convolutional counterparts on the one hand, and compared to models that lack symbolic grounding during training on the other hand. Our findings highlight the necessity to rethink the current intrinsic evaluations for acoustic word embeddings.

Familiar Words but Strange Voices: Modeling the Influence of Speech Variability on Word Recognition

Alexandra Mayn, Badr M. Abdullah, Dietrich Klakow

Summary by Badr M. Abdullah

In this work, we present a deep neural model of spoken word recognition which is trained to retrieve the meaning of a word (in the form of a word embedding) given its spoken form, a task that resembles that faced by a human listener. Our model is trained on naturalistic data that consists of actual acoustic realizations of word-aligned speech data extracted from the German portion of the Spoken Wikipedia Corpus. We then investigate the degree to which the emergent representations from the model can generalize with respect to two sources of variability in speech signals—inter-speaker variability and cross-lingual variability. Our experiments on speaker variability show the model is more sensitive to dialectical variation (when evaluated on the Swiss-German variety) than gender variation (when evaluated on a held-out gender). For the cross-lingual variability experiment, we evaluate the model on Dutch and English cognates in a zero-shot approach. We show that the model performs better at recognizing Dutch words compared to English words, which supports our intuition that cross-lingual word recognition performance should reflect language similarity.

Books

Finite-State Text Processing

By Kyle Gorman and Richard Sproat

Weighted finite-state transducers (WFSTs) are commonly used by engineers and computational linguists for processing and generating speech and text. This book first provides a detailed introduction to this formalism. It then introduces Pynini, a Python library for compiling finite-state grammars and for combining, optimizing, applying, and searching finite-state transducers. This book illustrates this library's conventions and use with a series of case studies. These include the compilation and application of context-dependent rewrite rules, the construction of morphological analyzers and generators, and text generation and processing applications.

Resources

EDICTOR. A Web-based Interactive Tool for Creating and Editing Etymological Datasets

By Johann-Mattis List

The EDICTOR is an interactive tool for creating, maintaining, and publishing etymological data which is stored in simple TSV format. After a longer time of silent development in which no official versions of the EDICTOR were published, a new version of the tool has now been officially published. This version is now available at the official EDICTOR website at https://digling.org/edictor/. The development version with functionalities to be published in future releases can be found at https://lingulist.de/edictor. Major features of the new EDICTOR version include: (1) a new panel allowing for the quick annotation of cognate morphemes by using cognate identifiers and morpheme glosses as well as the manual morpheme segmentation of words. (2) the indication of language subgroups, provided the data contains a column called "SUBGROUP" in the wordlist panel and in some of the alignment panels. (3) a new markup for cognate sets (mainly partial cognates), indicating whether they span a) several concepts, or are b) singletons, occuring only one time in the data. (4) the automated update of the data when working on multiple open instances for the same dataset with the help of the database version of the EDICTOR tool.

VoxLingua107: A Speech Dataset for Training Spoken Language Identification Models

By Tanel Alumäe and Jörgen Valk

VoxLingua107 is a large speech dataset for (pre-)training spoken language identification models, containing 6628 hours of speech from 107 different languages. The dataset was automatically collected from YouTube, by retrieving videos using language-specific search phrases (random trigrams from Wikipedia of the particular language). Although various post-processing steps were applied for filtering this data, there are still clips in the dataset that are not in the given language or contain non-speech. The average noise rate is around 2% overall, but it varies a lot according to the language.

A model trained on the dataset using the SpeechBrain toolkit is available here: https://huggingface.co/TalTechNLP/voxlingua107-epaca-tdnn.

Utterance embeddings averaged over the training data of the individual languages and projected into 2D space represent the structure of the language families remarkably well, see the paper describing the dataset for an interesting plot: https://arxiv.org/abs/2011.12998

Dataset URL: http://bark.phon.ioc.ee/voxlingua107/

Materials for Goals and Methods of Computational Linguistics

By Olga Zamaraeva, Julian Michael Emily Proch Ahn, Tsudoi Wada

Wonderful bibliography on the following topics:
History of computational linguistics (CL), CL today: Research questions, Evolution of methodologies, Evaluation, CL in Academy and Industry, The Future of CL

Discussions

What does “Native speaker” Mean, Anyway? (by Devin Grammon and Anna Babel)

Talks

SIGTYP Lectures

SIGTYP started its lecture series, our summer schedule is available here: https://sigtyp.io/lectures.html. The lectures are typically held on Fridays, we share a Zoom link via SIGTYP Google group, RocketChat channel, as well as via Eventbrite.

A talk typically consists of four ~15-min parts, each followed by a live Q&A session. The first part is an introduction to the main research topic, then the speaker discusses 2-3 research questions, followed by a conclusion and future directions.

Talks are then published on our Youtube channel.

SIGTYP on Youtube

SIGTYP now has a Youtube channel where we will be posting videos from our keynote speakers and other presenters who consent. This should allow us to reach a larger audience with our content! We

will also post recordings from the SIGTYP speaker series (held outside of regular conferences), which is still in development.

For our community in China: we created a channel on Bilibili https://space.bilibili.com/1055445444/channel/detail?cid=178350

SIGTYP 2021 -- All Talks, Papers, and Discussions are Publicly Available!

We would like to thank all who contributed to it! We invite everyone to provide some feedback to help us improve future events!

We are now planning to run a virtual SIGTYP2021 November Edition, please let us know if you would be willing to be a co-organizer (by sending an email to sigtyp@gmail.com).

Abralin ao Vivo – Linguists Online

Abralin ao Vivo – Linguists Online has a daily schedule of lectures and panel sessions with distinguished linguists from all over the world and from all subdisciplines. Most of the lectures and discussions will be in English. These activities will be broadcast online, on an open and interactive platform: abral.in/aovivo. The broadcasts will be freely available for later access on the platform afterwards.