SIGTYP

Recent Developments in Computational Typology and Multilingual Natural Language Processing

January 07, 2021 · Issue #8

Editors: Ekaterina Vylomova and Ryan Cotterell

This is SIGTYP's eighth 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.

Research Papers

Probing Multilingual BERT for Genetic and Typological Signals

Taraka Rama, Lisa Beinborn, Steffen Eger

Summary by Taraka Rama

We probe the layers in multilingual BERT (mBERT) for phylogenetic and geographic language signals across 100 languages and compute language distances based on the mBERT representations. We 1) employ the language distances to infer and evaluate language trees, finding that they are close to the reference family tree in terms of quartet tree distance, 2) perform distance matrix regression analysis, finding that the language distances can be best explained by phylogenetic and worst by structural factors, and 3) present a novel measure for measuring diachronic meaning stability (based on cross-lingual representation variability) which correlates significantly with published ranked lists based on linguistic approaches. Our results contribute to the nascent field of typological interpretability of cross-lingual text representations.

Mixed Evidence for Crosslinguistic Dependency Length Minimization

Zoey Liu

Summary by Zoey Liu

We investigate whether and to what extent the principle of Dependency Length Minimization (DLM) predicts crosslinguistic syntactic ordering preferences. More specifically, we ask: (i) is there a typological tendency for shorter constituents to appear closer to their syntactic heads in constructions with flexible constituent orderings? (ii) how does the extent of DLM in these constructions vary for languages with different structural characteristics? Our study uses prepositional and postpositional phrase (PP) typology as a testbed. Leveraging multilingual corpora for 34 languages, we focus on sentences with verb phrases that have exactly two PP dependents on the same side of the head verb, the ordering of which under certain conditions contains flexibility. Overall we show a pronounced preference for shorter PPs to be closer to the head verb, establishing the first large-scale quantitative evidence that DLM exists in crosslinguistic syntactic alternations. Furthermore, we present evidence that while the efficacy of DLM depends on the specific ordering structures of different language types, across languages there appears to be a much stronger preference for DLM when the two PPs appear postverbally, compared to no or a much weaker tendency for shorter dependencies when the two PPs occur preverbally. This contrast is the most visible in mixed-type languages with head-initial PPs that can appear both after or before the head verb. Within the limited number of rigid OV languages in our dataset, which have head-final PPs before the head verb, we observe no robust tendency for DLM, in contrast to the patterns in languages with head-initial PPs after the head verb. This contradicts previous findings of a longer-before-short preference in preverbal orders of head-final languages.

Re-evaluating Phoneme Frequencies

Jayden L. Macklin-Cordes and Erich R. Round

Summary by Jayden L. Macklin-Cordes and Erich R. Round

Causal processes can give rise to distinctive distributions in the linguistic variables that they affect. Consequently, a secure understanding of a variable's distribution can hold a key to understanding the forces that have causally shaped it. A storied distribution in linguistics has been Zipf's law, a kind of power law. In the wake of a major debate in the sciences around power-law hypotheses and the unreliability of earlier methods of evaluating them, here we re-evaluate the distributions claimed to characterize phoneme frequencies. We infer the fit of power laws and three alternative distributions to 166 Australian languages, using a maximum likelihood framework. We find evidence supporting earlier results, but also nuancing them and increasing our understanding of them. Most notably, phonemic inventories appear to have a Zipfian-like frequency structure among their most-frequent members (though perhaps also a lognormal structure) but a geometric (or exponential) structure among the least-frequent. We compare these new insights the kinds of causal processes that affect the evolution of phonemic inventories over time, and identify a potential account for why, despite there being an important role for phonetic substance in phonemic change, we could still expect inventories with highly diverse phonetic content to share similar distributions of phoneme frequencies. We conclude with priorities for future work in this promising program of research.

Consonant Co-occurrence Classes and the Feature-Economy Principle

Dmitry Nikolaev, Eitan Grossmann

Summary by Dmitry Nikolaev, Eitan Grossman

A central question in phonological typology (and in phonology more generally) is whether there are principles that govern the size, structure and constituent parts of phonological inventories, and if so, what they are. The feature-economy principle is one of the mainstays of contemporary discussions of phonological segment inventories in the languages of the world. Two different, albeit largely congruent, formulations of this principle were proposed by Lindblom & Maddieson ('small paradigms tend to exhibit 'unmarked' phonetics whereas large systems have 'marked' phonetics'; 1988) and Clements ('languages tend to maximise the ratio of sounds over features';2003).

In this paper, we test the explanatory power of this principle by conducting a study of the co-occurrence of consonant segments in phonological inventories, based on a sample of 2761 languages. We show that the feature economy principle is able to account for many important patterns in the structure of the world's phonological inventories; however, there are particular classes of sounds, such as what we term the 'basic consonant inventory' (the core cluster of segments found in the majority of the world's languages), as well as several more peripheral clusters whose organisation follows different principles.

The Impact of Information Structure on the Emergence of Differential Object Marking: an Experimental Study

Shira Tal, Kenny Smith, Jennifer Culbertson, Eitan Grossman, Inbal Arnon

Summary by Shira Tal et al.

Many languages exhibit differential object marking (DOM), where only certain types of grammatical objects are marked with morphological case. Traditionally, it has been claimed that DOM arises as a way to prevent ambiguity by marking objects that might otherwise be mistaken for subjects (e.g., animate objects). While some recent experimental work supports this account (Fedzechkina et al., 2012), research on language typology suggests at least one alternative hypothesis. In particular, DOM may instead arise as a way of marking objects that are atypical from the point of view of information structure. According to this account, rather than being marked to avoid ambiguity, objects are marked when they are given (already familiar in the discourse) rather than new. Here, we experimentally investigate this hypothesis using two artificial language learning experiments. We find that information structure impacts participants' object-marking, but in an indirect way: atypical information structure leads to a change of word order, which then triggers increased object marking. Interestingly, this staged process of change is compatible with documented cases of DOM emergence (Iemmolo, 2013). We argue that this process is driven by two cognitive tendencies. First, a tendency to place discourse given information before new information, and second, a tendency to mark non-canonical word order. Taken together, our findings provide corroborating evidence for the role of information structure in the emergence of DOM systems.

Using Lexical Language Models to Detect Borrowings in Monolingual Wordlists

John E. Miller, Tiago Tresoldi, Roberto Zariquiey, César A. Beltrán Castañón, Natalia Morozova, Johann-Mattis List

Summary by John E. Miller et al.

Lexical borrowing, the transfer of words from one language to another, is one of the most frequent processes in language evolution. In order to detect borrowings, linguists make use of various strategies, combining evidence from various sources. Despite the increasing popularity of computational approaches in comparative linguistics, automated approaches to lexical borrowing detection are still in their infancy, disregarding many aspects of the evidence that is routinely considered by human experts. One example for this kind of evidence are phonological and phonotactic clues that are especially useful for the detection of recent borrowings that have not yet been adapted to the structure of their recipient languages. In this study, we test how these clues can be exploited in automated frameworks for borrowing detection. By modeling phonology and phonotactics with the support of Support Vector Machines, Markov models, and recurrent neural networks, we propose a framework for the supervised detection of borrowings in mono-lingual wordlists. Based on a substantially revised dataset in which lexical borrowings have been thoroughly annotated for 41 different languages from different families, featuring a large typological diversity, we use these models to conduct a series of experiments to investigate their performance in mono-lingual borrowing detection. While the general results appear largely unsatisfying at a first glance, further tests show that the performance of our models improves with increasing amounts of attested borrowings and in those cases where most borrowings were introduced by one donor language alone. Our results show that phonological and phonotactic clues derived from monolingual language data alone are often not sufficient to detect borrowings when using them in isolation. Based on our detailed findings, however, we express hope that they could prove to be useful in integrated approaches that take multi-lingual information into account.

Multilingual Speech Translation with Efficient Fine-tuning of Pretrained Models

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, Michael Auli

Summary by Xian Li et al.

We present a simple yet effective approach to build multilingual speech-to-text (ST) translation by efficient transfer learning from pretrained speech encoder and text decoder. Our key finding is that a minimalistic LNA (LayerNorm and Attention) finetuning can achieve zero-shot crosslingual and cross-modality transfer ability by only fine-tuning less than 10% of the pretrained parameters. This enables effectively leveraging large pretrained models with low training cost. Using wav2vec 2.0 for acoustic modeling, and mBART for multilingual text generation, our approach advanced the new state-of-the-art for 34 translation directions (and surpassing cascaded ST for 23 of them) on large-scale multilingual ST benchmark CoVoST 2 (+6.4 BLEU on average across 15 En-X directions and +5.1 BLEU on average across 19 X-En directions). Our approach demonstrates strong zero-shot performance in a many-to-many multilingual model (+5.7 BLEU on average across 18 non-English directions), making it an appealing approach for attaining high-quality speech translation with improved parameter and data efficiency.

Geographical and Social Isolation Drive the Evolution of Austronesian Languages

Cecilia Padilla-Iglesias, Erik Gjesfjeld, Lucio Vinicius

Summary by Cecilia Padilla-Iglesias et al.

The origins of linguistic diversity remain controversial. Studies disagree on whether group features such as population size or social structure accelerate or decelerate linguistic differentiation. While some analyses of between-group factors highlight the role of geographical isolation and reduced linguistic exchange in differentiation, others suggest that linguistic divergence is driven primarily by warfare among neighbouring groups and the use of language as a marker of group identity. Here we provide the first integrated test of the effects of five historical sociodemographic and geographic variables on three measures of linguistic diversification among 50 Austronesian languages: rates of word gain, loss and overall lexical turnover. We control for their shared evolutionary histories through a time-calibrated phylogenetic sister-pairs approach. Results show that languages spoken in larger communities create new words at a faster pace. Within-group conflict promotes linguistic differentiation by increasing word loss, while warfare hinders linguistic differentiation by decreasing both rates of word gain and loss. Finally, we show that geographical isolation is a strong driver of lexical evolution mainly due to a considerable drift-driven acceleration in rates of word loss. We conclude that the motor of extreme linguistic diversity in Austronesia may have been the dispersal of populations across relatively isolated islands, favouring strong cultural ties amongst societies instead of warfare and cultural group marking.

Shared Tasks (SIGTYP 2020 -- Watch Online)

SIGTYP 2020 Shared Task: Prediction of Typological Features

Johannes Bjerva, Elizabeth Salesky, Sabrina J. Mielke, Aditi Chaudhary et al.

Summary by Johannes Bjerva

The SIGTYP 2020 Shared Task on prediction of typological features has now concluded! A big thanks to all five teams who participated, and the co-organisers who made this happen. Participating teams built systems for predicting typological features in WALS, with an evaluation set-up including controls for both phylogenetic relationships and geographic proximity. The evaluation focussed on 6 genera: Tucanoan, Madang, Mahakiranti, Nilotic, Mayan, and Northern Pama-Nyungan. These were chosen so as to give us one per macro-area in WALS, thus yielding a considerable typological spread across the world. This resulted in a challenging setting where the best system (ÚFAL, Vastl et al. (2020)) obtained a macro-averaged accuracy of 75%. The most difficult features to predict were, unsurprisingly, the ones which were the least frequent in the training data. Nonetheless, the top four systems achieved >65% accuracy on these features, whereas other systems achieved ~20% accuracy on these.

You may watch the share task presentations online:

1. SIGTYP 2020 Shared Task Overview by Johannes Bjerva - https://slideslive.com/38939790
2. ÚFAL team: Predicting Typological Features in WALS using Language Embeddings and Conditional Probabilities - https://slideslive.com/38939792
3. NEMO: Frequentist Inference Approach to Constrained Linguistic Typology Feature Prediction - https://slideslive.com/38939791
4. KMI-Panlingua-IITKGP: Exploring Rules and Hybrid Systems for Automatic Prediction of Typological Features - https://slideslive.com/38939795
5. Gerhard Jäger: Imputing Typological Values via Phylogenetic Inference - https://slideslive.com/38939793
6. NUIG: Multitasking Self-Attention based Approach - https://slideslive.com/38939794

Resources

COLING Tutorial “Cross-lingual Semantic Representation for NLP with UCCA”

Taught by Omri Abend, Dotan Dvir, Daniel Hershcovich, Jakob Prange and Nathan Schneider

This is an introductory tutorial to UCCA (Universal Conceptual Cognitive Annotation), a cross-linguistically applicable framework for semantic representation, with corpora annotated in English, German and French, and ongoing annotation in Russian and Hebrew. UCCA builds on extensive typological work and supports rapid annotation. The tutorial provides a detailed introduction to the UCCA annotation guidelines, design philosophy and the available resources; and a comparison to other meaning representations. It also surveys the existing parsing work, including the findings of three recent shared tasks, in SemEval and CoNLL, that addressed UCCA parsing. Finally, the tutorial presents recent applications and extensions to the scheme, demonstrating its value for natural language processing in a range of languages and domains.

CMU Course "Multilingual Natural Language Processing"

Taught by Graham Neubig, Yulia Tsvetkov, Alan Black. Lecture videos, slides, and homework assignments will be publicly available.

Students who take this course should be able to develop linguistically motivated solutions to core and applied NLP tasks for any language. This includes understanding and mitigating the difficulties posed by lack of data in low-resourced languages or language varieties, and the necessity to model particular properties of the language of interest such as complex morphology or syntax. The course introduces modeling solutions to these issues such as multilingual or cross-lingual methods, linguistically informed NLP models, and methods for effectively bootstrapping systems with limited data or human intervention. The project work involves building an end-to-end NLP pipeline in a language you don't know.

BiValTyp - A Database of Bivalent Verbs

BivalTyp is a typological database of bivalent verbs and their encoding frames. As of 2020, the database presents data for 85 languages, mainly spoken in Northern Eurasia. The database is based on a questionnaire containing 130 predicates given in context. Language-particular encoding frames are identified based on the devices (such as cases, adpositions, and verbal indices) involved in encoding two predefined arguments of each predicate (e.g. 'Peter' and 'the dog' in 'Peter is afraid of the dog'). In each language, one class of verbs is identified as transitive. The goal of the project is to explore the ways in which bivalent verbs can be split between the transitive and different intransitive valency classes.

Talks

SIGTYP 2020 Keynote Talks -- Watch Online

1. Richard Sproat: Taxonomy of Writing Systems: How to Measure How Logographic a System is. Slideslive: https://slideslive.com/38939787
2. Miriam Butt: Building Resources: Language Comparison and Analysis. Slideslive: https://slideslive.com/38939785
3. Harald Hammarström: How Many Languages are There in the World? Slideslive: https://slideslive.com/38939789
4. Yulia Tsvetkov: Speaker-centered NLP with Finer-Grained Multilingual Model Proficiency. Slideslive: https://slideslive.com/38939786
5. Bill Croft: Crosslingual Syntactic and Semantic Annotation: Typological and Ethical Precepts. Slideslive: https://slideslive.com/38939788

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 aftewards.

Free Online course on East Caucasian languages

The East Caucasian languages form a deep-level family that is considered indigenous to the Caucasus. It consists of 30-50 distinct languages (according to different classifications), and is in fact largely responsible for the high rate of language density that the Caucasus as a linguistic area is famous for. The languages of the family feature a number of striking features, including rich consonant inventories, pervasive gender agreement with unusual targets, and complex systems of nominal spatial inflection, among other traits.