Recent Developments in Computational Typology and Multilingual Natural Language Processing

September 7, 2020 · Issue #6

Editors: Ekaterina Vylomova and Ryan Cotterell

This is SIGTYP's sixth 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 Taraka Rama, Simon Greenhill, Limor Raviv, Piotr Żelasko, Marieke Schouwstra describe their recent publications on linguistic typology and multilingual NLP.

Research Papers

A Cross-Linguistic Comparison of the Evolved Complexity of Numeral Systems

Simon Greenhill, Sieghard Beller, Andrea Bender, Hans-Jörg Bibiko, Eugene Chan, Robert Forkel, Russell Gray, Fiona Jordan, Christoph Rzymski, and Annemarie Verkerk

Summary by Simon Greenhill

The ways in which languages keep track of quantities differ substantially across the languages of the world (Bender & Beller, 2018). However, the global diversity of these systems has barely been explored.

Here we present Numeralbank: a new global database of numeral systems containing ~186,000 number words from ~5300 languages. Numeralbank contains number words for basic numbers from 1 to 10, and higher numbers (20, 30, etc), and regionally important numbers like 46,656 (6 to the power of 6, encountered southern New Guinea when counting piles of yams as a trade good).

In our preliminary investigations of this database, we show that, first, there is a strong relationship between a number and the orthographic length of its lexeme, where the lexical forms for numbers below five are shortest, followed by the numbers below ten. This pattern presumably follows from the fact the smaller numbers are often recruited into compositional systems to create larger numerals by combining smaller bases. Numbers that are not recruited as bases often (7, 8, 9) tend to be longer. Further, number words for multiples of base 10 (e.g. “twenty”) also tend to be short.

Second, we develop a novel method for characterizing the complexity of numeral systems in these languages, and quantify and model their evolution over time.
Our preliminary results show that languages like Mandarin, Tagalog and German have more efficient counting systems than most other languages.

Finally, we use these data to test some broad-scale generalizations about the dynamics and evolution of numeral base systems. Our findings suggest that, for example, base 4 or base 6 counting systems are largely unstable and languages evolve away from these states rapidly. In contrast, base 20 systems are very stable and attractive, especially if the language starts with a base 4 or 8 system. Decimal systems are universally attractive.

Work on Numeralbank is currently ongoing, and we are aiming for release towards the end of the year.

The Role of Community Size and Social Network Structure in the Emergence of Linguistic Structure

Limor Raviv, Antje Meyer, Shiri Lev-Ari

Summary by Limor Raviv

Why are there so many different languages in the world? How much do languages differ from each other in terms of their linguistic structure? And how do such differences come about?

One possibility is that linguistic diversity stems from differences in the social environments in which languages evolve. Specifically, it has been suggested that small, tightly knit communities can maintain high levels of linguistic complexity, while bigger and sparser communities tend to have languages that are structurally simpler, i.e., languages with more regular and more systematic grammars.

However, to date this hypothesis has not been tested experimentally. Moreover, community size and network structure are typically confounded in the real-world, making it hard to evaluate the unique contribution of each social factor to this pattern of variation.

To address this issue, we used a novel group communication paradigm. This experimental paradigm allowed us to look at the live formation of new languages that were created in the lab by different micro-societies under different social conditions. By analyzing the emerging languages, we could tease apart the causal role of community size [1] and network structure [2], and see how the process of language evolution and change is shaped by the fact that languages develop in communities of different sizes and different social structures.

In the first paper, we show that larger groups created languages with more systematic grammars, and did so faster and more consistently than small groups [1]. This finding suggested that social environment in which languages evolve, and specifically the number of people in the community, can affect the grammar of languages. We suggest that larger groups are under a stronger pressure to create systematic languages, given the fact that members of larger groups are typically faced with more input variability, and have less shared history with each member of their group.

In contrast, in the second paper we found no evidence for a similar role of network connectivity: all groups developed languages that were highly systematic, communicatively efficient, stable, and shared across members, with dense and sparse groups reaching similar levels of linguistic structure over time [2]. Although there were no significant differences between networks with respect to their degree of systematic grammar, we found that small-world networks showed the most

variance in their behaviors. This suggests that small-world networks may be more sensitive to random events (i.e., drift).

More: [1] Raviv, L., Meyer, A., Lev-Ari, S. (2019b). Larger communities create more systematic languages. Proceedings of the Royal Society B: Biological Science, 286(1907). doi:10.1098/rspb.2019.1262

[2] Raviv, L., Meyer, A., & Lev-Ari, S. (2020). The role of social network structure in the emergence of linguistic structure. Cognitive Science, 44(8), e12876. doi:10.1111/cogs.12876

That Sounds Familiar: an Analysis of Phonetic Representations Transfer Across Languages

Piotr Żelasko, Laureano Moro-Velázquez, Mark Hasegawa-Johnson, Odette Scharenborg, Najim Dehak

Summary by Piotr Żelasko

Only a handful of the world's languages are abundant with the resources that enable applications of speech processing technologies. One way to overcome this problem is to use the resources existing in other languages to train a multilingual automatic speech recognition (ASR) model, which, intuitively, should learn some universal phonetic representations. In this work, we focus on gaining a deeper understanding of how general these representations might be and how individual phones are getting improved in a multilingual setting.

To that end, we select a phonetically diverse set of thirteen languages and perform a series of monolingual, multilingual, and crosslingual (zero-shot) experiments. The ASR system is an end-to-end Transformer model, trained to recognize the International Phonetic Alphabet (IPA) token sequences with joint CTC and attention decoders. In particular, we ask the following questions: Can language-unique phones benefit from multilingual training? Does a phone improve more in a multilingual system when it occurs in more languages? Do phones shared by more languages transfer representations better in a cross-lingual scenario? How do the manner and the place of articulation affect the performance in cross-lingual and multilingual schemes? Are there phones with universal representations in our model?

We attempt to answer these questions with a detailed analysis of the model's errors. We observe significant improvements across all languages in the multilingual setting, and stark degradation in the crosslingual setting, where the model, among other errors, considers Javanese as a tone language. Notably, as little as 10 hours of the target language training data tremendously reduce ASR error rates. Our analysis uncovered that even the phones that are unique to a single language

could benefit greatly from adding training data from other languages - an encouraging result for the low-resource speech community.

From the World to Word order: Deriving Biases in Noun Phrase Order from Statistical Properties of the World

Jennifer Culbertson, Marieke Schouwstra, Simon Kirby

Summary by Marieke Schouwstra

Why are human languages structured the way they are? On the one hand, languages show striking diversity, but on the other hand, we can observe patterns that suggest this diversity is shaped by human cognition. In this paper we will focus on the structure of complex noun phrases, such as 'those two purple chairs'. The noun and the words that modify it can be ordered in 24 possible different ways, but only a subset of these orderings is represented in the languages of the world.

These preferred orderings have in common that the adjective always comes closer to the noun than the numeral, which in turn is closer to the noun than the demonstrative (this is an underlying structure). Orderings that violate this underlying structure can be found (for instance, Noun-Demonstrative-Numeral-Adjective, which is observed in Kîîtharaka), but only in a tiny proportion of the languages of the world. Almost all other languages are the result of transparent mappings between the underlying structure and the surface structure of the noun phrase.

We tested the hypothesis that there is a cognitive bias in favour of making such transparent mappings, by asking participants to improvise descriptions of complex noun phrases, by using only gesture and no speech. This method has been used to uncover cognitive biases driving constituent ordering, while limiting the influence of the native language of the participants. Our participants (N=20, all native speakers of English) showed a strong preference in favour of transparent mappings, while not producing any English orders.

To investigate where the underlying structure that drives noun phrase ordering might come from, we conducted a corpus study. We hypothesised that the underlying structure might be learnable by observing the world. Many objects in the world are closely associated with their properties, for instance, apples with the colours red and green, and wine with red or white. By contrast, it's less likely for objects to be associated closely with their numerosities, and even less so with their discourse status.

We used linguistic corpora as a proxy for the world, and measured the strength of association between the noun and each of the modifiers adjective, numeral, and demonstrative in data from 24

different languages (plus the English corpora in CHILDES), and confirmed the hypothesised pattern in each.

These results together suggest that our experience with objects in the world, combined with a preference for transparent mappings from conceptual structure to linear order, can explain constraints on noun phrase order in the languages of the world.

A Test of Generalized Bayesian Dating: A New Linguistic Dating Method

Taraka Rama and Søren Wichmann

Summary by Taraka Rama

In current practice, when dating the root of a Bayesian language phylogeny the researcher is required to supply some of the information beforehand, including a distribution of root ages and dates for some nodes serving as calibration points. In addition to the potential subjectivity that this leaves room for, the problem arises that for many of the language families of the world there are no available internal calibration points. Here we address the following questions: Can a new Bayesian framework which overcomes these problems be introduced and how well does it perform? The new framework that we present is generalized in the sense that no family-specific priors or calibration points are needed. We moreover introduce a way to overcome another potential source of subjectivity in Bayesian tree inference as commonly practiced, namely that of manual cognate identification; instead, we apply an automated approach. Dates are obtained by fitting a Gamma regression model to tree lengths and known time depths for 30 phylogenetically independent calibration points. This model is used to predict the time depths of both the root and the internal nodes for 116 language families, producing a total of 1,287 dates for families and subgroups. It turns out that results are similar to those of published Bayesian studies of individual language families. The performance of the method is compared to automated glottochronology, which is an update of the classical method of Swadesh drawing upon automated cognate recognition and a new formula for deriving a time depth from percentages of shared cognates. It is also compared to a third dating method, that of the Automated Similarity Judgment Program (ASJP). In terms of errors and correlations with known dates, ASJP works better than the new method and both work better than automated glottochronology.

Blogs

The Key to Language is Universal Psychology, Not Universal Grammar

A book summary by Paul Ibbotson and Misha Ketchell

Noam Chomsky argues that language gets its own ring-fenced mental processor, areas of which cannot be accessed by other non-linguistic aspects of cognition. This mental module comes with innate content organised before we experience the world, designed to work exclusively on language.

The author makes the argument that over 50 years' worth of developmental psychology and psycholinguistic research now demonstrates how the mental modular view vastly underestimates both the breadth and depth with which cognition interacts with, constrains and predicts language use. This interaction penetrates so deeply that it's reasonable to claim that language is built out of this general purpose psychological toolkit – abilities like memory, attention, inhibition, categorisation and social-cognition (that related to our social interactions) – rather than the language-specific one Chomsky and others had in mind.

Why the Linguist Needs the Historian

By James McElvenny

A fascinating turn in recent Natural Semantic Metalanguage (NSM) scholarship is the development of 'Minimal English', an international auxiliary language combining the best of Standard English and NSM. One of the earliest published mentions of this project is in Anna Wierzbicka's 2014 Imprisoned in English, where she states that Minimal English 'is, essentially, the English version of "Basic Human"', the rendering in English exponents of the set of primitive concepts uncovered by NSM research.

In the most detailed published treatment of Minimal English to date, the 2018 edited volume Minimal English for a Global World, Cliff Goddard and Anna Wierzbicka dedicate a section of their co-authored chapter to comparing Minimal English and Basic English.

However, the comparison misses many significant points of contact between the two endeavours.

In terms of the differences in 'structure', Goddard and Wierzbicka point out that Ogden's core vocabulary of 850 'Basic Words' does not respect the cross-linguistic primitives proposed within the NSM framework. In addition, a central grammatical feature of Basic English was the elimination of verbs from the language, a goal foreign to NSM thinking. These two differences between Basic and Minimal English are quite real, but focusing on them misses the more profound philosophical and methodological similarity between the projects: that both NSM and Basic English are centred around reductive paraphrase.

Resources

Atlas of Endangered Alphabets

Indigenous and minority writing systems, and the people who are trying to save them.
Watch an introduction to the project online: https://www.youtube.com/watch?v=vGLqLQ2pc00

Note: The Endangered Alphabets Project needs two interns for the fall. Ideal for a grad student in linguistics. Contact: Tim Brookes <tim@endangeredalphabets.com>

CMU Course "Multilingual Natural Language Processing"

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

Interslavic Language Project -- Projekt Medžuslovjanskogo Jezyka

The Interslavic language has been created by an author group on the basis of older projects: Slovianski from 2006 and Neoslavonic from 2009. The Interslavic language is at the very centre of the modern Slavic languages which all Slavs can understand without any prior study and use after some minimal learning only.

Talks

Abralin ao Vivo – Linguists Online

Abralin ao Vivo – Linguists Online has a daily schedule of lectures and panel session 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.

Dynamics of Language

Corpus annotation for typological research in discourse and grammar: the Multi-CAST initiative
Stefan Schnell, University of Bamberg, Friday 11 September, 4:00pm AEST RSVP for zoom link t.barratt@unimelb.edu.au

How Language Shapes Thought: Categorisation in the South Pacific

We invite you to discover languages embedded in a distant culture, which offer a window into human cognition.

We discuss several endangered indigenous languages in the Pacific Island countries of Vanuatu and New Caledonia. Each language has a set of classifiers, linguistic markers that categorise a person's possessions into different groups based on how they intend to use them. For example, speakers have to say 'I will drink my drinkable coconut' or 'eat my edible coconut' depending on how they intend to use their coconut. Different languages in the South Pacific have different numbers of classifiers ranging from two to well over twenty.
Date: Thursday 12 November, 01:00 – 02:00 AEDT