Abstracts

Willem F.H. Adelaar (Universiteit Leiden, Centre for Linguistics)

On the status of verbal aspect in Central Peruvian Quechua

In recent typological literature on verbal morphology Aspect is often presented as closely interrelated with Tense and Mood (TAM). The Quechuan languages provide a clear illustration that Tense and Aspect at the least can be analyzed as mutually independent categories which occupy different locations and play different roles in an otherwise rigidly organized verbal structure. Whereas Tense distinctions are a general feature of the Quechuan language family, Aspect appears to be a relatively recent addition to Quechuan verbal morphology. It has attained a high level of complexity in the Central Peruvian branch of these languages generally known as Quechua I (Torero 1964).

From a structural point of view, Central Peruvian Aspect markers may merge with Plural number markers, which are also of recent formation. Meanwhile, there is no sign of formal fusion between Aspect and Tense, which continue to operate independently from each other. A further significant and exclusive feature of Aspect in Central Peruvian Quechua is its incompatibility with the category of Negation.

Michael Cysouw (Philipps-Universität Marburg)

Dynamic universals in the linguistic marking of location

Using a massively parallel corpus (about 1500 Bible translations) I will show what is possible by looking at just one word across all those languages, viz. the word “Jerusalem”. Only using the written variation and without any need for annotation it is possible to replicate various well-known universals, but also to extend them into a diachronic ‘dynamic’ direction (cf. Cysouw 2011).


Hans Götzsche (Aalborg Universitet)

The issue of universals illustrated by analytic problems in phonology and syntax

On the background of the theme ‘linguistic universals and language typology’ I shall offer a presentation of

- some basic theoretical and philosophical problems when talking about ‘universals’,
- an approach to, specifically, syntactic theory that makes a clear distinction between the mental structures (the cognitive system(s)) that process sentences (clauses) and the formal system that has been developed to describe such structures.

The first part will take its point of departure the question of what a phoneme is. In his classical book Daniel Jones says that

‘... a phoneme is a FAMILY OF SOUNDS IN A GIVEN LANGUAGE WHICH ARE RELATED IN CHARACTER AND ARE USED IN SUCH A WAY THAT NO ONE MEMBER EVER OCCURS IN A WORD IN THE SAME PHONETIC CONTEXT AS ANY OTHER MEMBER. (Jones 1950/67 p. 10)

Apart from the well-known problem of the existence of words (see Haspelmath 2011) one may justifiably ask (i) what ‘a family of sounds’ is, (ii) what constitutes a ‘language’ as a uniform entity, and, (iii) finally, what “abstract” sounds (p. 6) are. So, when phonematic schemata obviously differ between languages and dialects one may ask: is, then, the category PHONEME as such a universal one? Especially when the features of the acoustic signal and the phonetic input that is filtered by the auditive system sometimes do not reveal any similarity with the alleged chain of phonemes. Later developments do not seem to have solved the problem, and the presentation will offer a number of ways to understand the widespread use of the theoretical notion. I will also indicate how copus based research may undermine the notion of phoneme and underpin alternative training methods in reading.

The second part will summarise and update the approach in formal syntax presented in Götzsche (2013b), the so-called Formative Grammar (FoG) theory. There are, and have been, a number of formal models on the market in linguistics, but this approach is based on a special kind of logic, called Occurrence Logic, that does not work on the basis of truth and time but instead on whether an entity, an item, can be said to occur or not. The FoG, then, is a combinatorial system that is used to describe actual sentences as structures of occurrence-interrelated units. The theory has been applied to Danish and Swedish, and I am now in contact with a high-level publisher concerning a ‘Formative Grammar of the English Language’. The theory is the only integration of Hjelmlevian Glossematics and the topological (so-called sentence table) syntactic theory developed in the first half of the 20th century by the Danish linguist Paul Diderichsen. The universality of the theory lies in the assumption that there are, in fact, universal syntactic categories, and this stance will be defended theoretically and empirically, partly by presenting a model of a cognitive system and partly by offering a number of analyses.

References:
Corinna Handschuh (Universität Regensburg)

Lexical and argument structure of naming constructions

Naming constructions in the languages of the world differ to a great extent with respect to the realization of argument/non-argument roles and the lexical structure of their predicates. For this study, I use various sources (language experts, corpus data and the ValPaL-database, Hartmann et al. 2013) to investigate the constructions employed to express the following contexts cross-linguistically:

(1) The parents named/called the child Anna.
(2) The child is called Anna.

Of particular interest is the status of the NAME in construction (1). At first glance one might consider it to be a normal di-transitive construction in English. Yet on closer inspection, it becomes apparent that the NAME does not share the properties of English secondary objects, e.g. taking part in the Dative-Alternation. Matushansky (2008) argues that, in a number of languages including English, the construction constitutes a small clause in which the NAMEE is the subject and the NAME the predicate. Anderson (2007:219) on the other hand considers the NAME to be neither a complement nor to be predicative. The most promising analysis, in my opinion, is that it is an instance of mentioning the name without a referential use of it. In other words, the form of the name can most accurately be described as a quotation or citation. Indeed, a number of languages in my current dataset have an overt quotative marker use it on the name in such constructions. However, other languages, e.g. Igbo and Teop, show evidence that the name has argument status in these constructions.

Another interesting point related to naming-constructions is the variation between languages in how the predicates of the construction (1) and (2) are related to one another. So far I have discovered the following options: different verb stems for (1) and (2), e.g. German *nennen* vs. *heißen*; non-verbal constructions for (2), e.g. the type “Xs name is Y”; light verb constructions for (1), e.g. “give name Y to X”; valency-changing operations between (1) and (2). While valency changes can logically happen in both directions – increasing the valency of (2) to form (1) or decreasing the valency of (1) to form (2), e.g. by passivization – the second option appear to be much more common in the languages I could investigate so far. One example of the – supposedly rare – valency-increasing construction is provided by East Futunan (Claire Moyse-Faurie, personal communication), in which *igoa* ‘a name/be named’ has a causativised counterpart *faka’igoa* ‘give a name to s.o.’.

In this paper, I will give a cross-linguistic overview of the structure naming-constructions with respect to the status/encoding of the NAME and the lexical relationship between the predicates in the different constructions. Based on this example, I will also address methodological issues in typological research, discuss the strengths and weaknesses of different approaches to data collection, and hopefully have a fruitful discussion on how different kinds of data can be combined in research in order to broaden our understanding of linguistic diversity.

References:

**Martin Haspelmath (Universität Leipzig & Max-Planck-Institut für Menschheitsgeschichte)**

Differential place marking and differential object marking

This paper gives an overview of differential place marking phenomena and formulates a number of universals that seem to be well supported. Differential place marking is a situation in which the coding of locative, allative or ablative roles depends on subclasses of nouns, in particular toponyms, inanimate common nouns and human nouns. When languages show asymmetric coding differences depending on such subclasses, they show shorter (and often zero) coding of place roles with toponyms, and longer (often adpositional rather than affixal) coding of place roles with human nouns. Like differential object marking, differential place marking can be explained by frequency asymmetries, expectations derived from frequencies, and the general preference for efficient coding. I also argue that differential place marking patterns provide an argument against the need to appeal to ambiguity avoidance to explain differential object marking.

**Nikolaus P. Himmelmann (Universität zu Köln)**

Nouns and Verbs

In some languages, nouns are clearly distinguished from verbs with regard to their morphosyntactic and morphological properties, in others they are not. For a few languages (language groups), it has in fact been claimed that no (grammatical) distinction between nouns and verbs exists. The controversy regarding the latter claim is clouded by a number of issues (e.g. the failure to properly distinguish between syntactic and lexical categories). But, more importantly perhaps, it has obscured the uncontroversial fact that there is indeed a great range of variation with regard to the kind and amount of grammatical differences between nouns and verbs.

Among a number of issues pertaining to the variable strength of the noun/verb distinction, it is unclear to date whether the cross-linguistically observable variation in this regard has consequences that go beyond the domain of the purely grammatical (i.e. consequences for grammatical description and challenges for grammatical theory).

The talk will briefly illustrate the possibility of such non-grammatical correlates with regard to issues of neurocognitive processing (are verbs processed differently from nouns?) and usage differences that go beyond the narrowly grammatical, in particular frequency distributions in natural discourse.
Natalia Levshina (Universität Leipzig)

Token-based typology: word order in the Universal Dependencies corpora

The present paper discusses the benefits and challenges of token-based typology, which takes into account the frequencies of words and constructions in discourse. First, this approach makes it possible to introduce new criteria for language classification and formulate new cross-linguistic generalizations, which would be difficult to achieve with the traditional, type-based approach. This point is illustrated by a case study of word order flexibility, which demonstrates a non-linear relationship between word order entropy and constituent length. Second, it allows us to provide an integrative account of categorical and gradable phenomena which are shaped by the same cognitive and functional pressures. As an example, I provide a corpus-based study of heaviness hierarchy of noun modifiers. Third, the token-based approach helps us to capture and analyse the subtle differences and similarities between closely related languages, as illustrated by a glottometric case study of word order. The case studies are based on data from the Universal Dependencies corpora.

Susanne Michaelis (Universität Leipzig & Max-Planck-Institut für Menschheitsgeschichte)

Asymmetry in tense-aspect coding: Creole data support a universal trend

In quite a few creole languages, stative verbs are zero-marked for present-time reference (see ex. 1a), whereas dynamic verbs show overt marking for the same time reference (see ex. 1b).
(1) Haitian Creole (DeGraff 2007)

- Bouki Ø renmen chat la. Bouki love cat DEF'Bouki loves the cat'.
- M ap manje. 1SG IPFV eat
- 'I am eating'.

Asymmetric coding is also found for past-time reference. But this time, stative verbs are overtly marked (see ex. 2a) while dynamic verbs are zero-marked (see ex. 2b).
(2) Haitian Creole

- Bouki té konn repons lan. Bouki PFV know answerDEF'Bouki knew the answer'.
- Bouki Ø vann chat la. Bouki sell cat DEF'Bouki sold the cat'.

What we see here is a skewed distribution of zero and overt markers as the result of an interaction between actionality classes (stative/dynamic verbs, or more broadly situations) and tense-aspect coding (imperfective/perfective) to refer to either present time or to past time. Since corpus data seem to show that stative verbs more frequently occur with present-time reference ('I love'), and dynamic verbs more frequently occur with non-present-time reference, in English with Past Tense ('I ate'), this coding asymmetry is a kind of grammatical form– frequency correspondence (Haspelmath et al. 2014).
This means that the rarer combinations of actionality classes and aspect classes are coded longer than or equally long as the more frequent combinations: here ap, té in 1b and 2a vs. zero-coding in 1a and 2b. Such a coding asymmetry can be seen as a functional response to the need to highlight rarer, less predictable actionality/aspect combinations.

This tendency for asymmetric coding means that when there is an asymmetry, it generally goes in the same direction, not that there must be an asymmetry, or even that there tends to be an asymmetry. Indeed, cross-linguistically the actionality/aspect coding asymmetry does not seem to be widely distributed (Šluinskij 2012), and we see much more symmetric coding. But most importantly, we do not see counter-asymmetric patterns: longer coding for the frequent combinations, and zero coding for the rarer combinations.

In this talk, I will show that data from high-contact languages (pidgins and creoles) support this universal trend (see Michaelis et al. 2013, Maurer & APiCS Consortium 2013). More than half of the 76 pidgins/creoles in APiCS from different parts of the world show a pattern similar to Haitian Creole, displaying the coding asymmetry just described. Other creoles feature symmetrical coding in stative/dynamic verbs with respect to present-time and past-time reference. But crucially we do not find counter-examples to the universal coding asymmetry.

In traditional creole linguistics, these asymmetries have been explained by a bioprogramme (Bickerton 1981) or by substrate influences (e.g. Lefebvre 1998), but these functional considerations show that neither is strictly necessary.

References:

Olav Mueller-Reichau (Universität Leipzig)

Shorter and longer forms in Russian secondary imperfectives

Russian aspect is a classificatory category (Plungjan 2011; Arkadiev & S’luinskij 2015) with each verbal stem being associated with one out of two grammatical values, perfective or imperfective.
The assignment of aspect to stems is governed by two default rules (1), whose output may be overwritten by additional overt morphology (Bohnemeyer & Swift 2004; Lehmann 2010; Mende et al. 2012): by the suffix yva- (manifesting itself in different allomorphs) or by so-called “selectionally restricted” (Tatevosov 2013) prefixes (2).

1. telic stem → pf 
2. atelic stem → ipf

(2) i. telic stem + -yva → ipf
ii. prefixsel. restr. + atelic stem → pf

The overall coding of Russian verbs meets the predictions of the form-frequency correspondence principle FFC argued for in, e.g. Haspelmath (2014); Haspelmath et al. (2014): Out of 50,214,412 verb tokens included in the Russian National Corpus RNC, 28,563,260 are imperfective, and out of the imperfectives, only 2,697,259 verb forms contain suffix -yva. Thus, roughly half of all verb forms are simplex imperfectives (in the absence of -yva, prefixed verbs are perfective). If scope is limited to verbs describing telic event types, we observe “markedness reversal”: 2,697,259 telic imperfectives of longer form (those imperfectives formed via (2i), called “secondary imperfectives”) oppose to 23,358,743 perfectives of shorter form (but note that one would still have to subtract those perfectives which are formed via (2ii) from the latter number).

Interestingly from the perspective of FFC, Modern Russian displays doublet forms for secondary imperfectives SI within the class of perfective verbs with theme vowel /i/ (Zaliznjak & Smelev 1997:60). An example is shown in (3):

(3) suzit' pf ↔ sužat' ipf suživat' ipf (‘to narrow’)

In these cases, there is a shorter and a longer SI available. “Shorter SIs” equal perfectives in syllable number in those forms that are built on the basis of the infinitive stem, but display one more syllable in forms built from present tense stems. “Longer SIs” are longer in syllable number throughout the verbal paradigm. Zaliznjak & Smelev report that shorter SIs are successively being replaced by longer SIs. Since this development strengthens asymmetrical coding, it may be viewed as an adaptation to FFC.

To get a clearer picture, I have checked for the token frequencies of the doublet forms in the mass media subcorpus of RNC, limited to text documents which have been published from 2000 until 2018 (comprising 228,521,421 words; 14/07/2018). The results are summarized in Figure 1; the verbs chosen are those mentioned in Zaliznjak & Smelev (1997).
The counts marked by a star are meaningless for the present purposes because they include the opposite SI-forms (in these cases, the doublets appear to be subsumed under one lemma in the corpus). To get an impression of the distribution of shorter and longer SIs also for these cases, I ran another query counting particular verb forms of the values infinitive, third person singular and past tense masculine (using the same subcorpus), see Figure 2.
We observe that SI doublets fall into two groups. For the first group of verbs, the expansion of longer imperfectives has advanced far already, sometimes carried out completely. Not so for the members of the second group, which in our sample are prostudit’sja, podmenit’, suzit’ and postanovit’. These verbs are obviously much more (and sometimes even perfectly) reluctant to lengthening their SIs.

My explanation for this reluctance is rooted in the semantics of the respective verbal lexemes (that morphology cannot be the critical factor can be seen from that postanovit’ behaves different from, e.g., ustanovit’). The generalization is: For a SI to adopt the longer form, the event type described by the verbal predicate must be such that it is realizable in a step-by-step manner, i.e. as a serialization of subevents, involving either repetitions of the same action or sequences of different actions that on the whole make up the (macro) event described by the predicate.
Where this condition is not naturally met, as with the meanings ‘to catch a cold’, ‘to substitute’, ‘to narrow’ and ‘to decide’, the shorter forms will be the preferred ones. The reason for the dispreference of the longer form is, arguably, that the suffix -yv-/iv- still carries the information of iterativity, which it displayed productively in the 16th to 18th centuries (Isaenko 1968:367; Paduˇceva 2015), when forming iterative stems from simple stems denoting atelic processes (e.g. pisat’ ‘write’) → pisyvat’ ‘write from time to time’). The option to optimize FFC is dismissed due to a semantic mismatch between the lexeme meaning and the content of the additionally morpheme in the longer form.

References:

Michael Richter, Gerhard Heyer (Universität Leipzig)

Interaction of Information Content and Frequency as predictors of verbs’ lengths

The aim of our study is to test whether verbs’ lengths are contingent on their Entropy, the research question being: How do the predictors Information Content and Frequency interact in a linear model when predicting aspectual verb forms? The verbs occur either in Perfective or Imperfective aspect. 30 Languages are in the focus of our study, data resources are UD treebanks (version 2.1).
Frequency is the classical Zipfian predictor and expresses the idea that The magnitude of words tends, on the whole, to stand in an inverse [...] relationship to the number of occurrences (Zipf 1936). Information Content (IC, Cohen Priva, 2008, Piantadosi, 2011) represents the intuition that longer words should carry more information than shorter words, otherwise the greater length would be not economical. We estimated Information Content from n-grams to the left of our targets.

The results of our study indicate (i) that linear models employing the predictors IC and Frequency yield high correlations between predicted and actual verbs’ lengths and (ii) that IC and Frequency do not have a simultaneous positive impact on the length of verb forms. IC is just as large as it needs to be in order to predict verb lengths. The IC-effect is high when Frequency has low values, i.e. with frequent verb forms. The length of rare verb forms is predicted by Frequency and in these cases the effect of IC is almost annihilated.

References:

Rebecca Voll (Universität Leipzig)
Corpora of Minority Languages and Endangered Languages - an Inventory for Corpus Linguists.

This presentation is meant (i) to call the attention of corpus linguists to endangered languages and minority languages corpora, (ii) to explore the synergy potential of these two branches of linguistics, and (iii) to serve as a rough practical guide to their use and usefulness, specifically tailored to corpus linguists. In the last years, various authors (such as McEnery & Ostler 2000, Johnson 2004 and Cox 2011) have encouraged the collaboration of documentary linguistics and corpus linguistics. They have listed possible uses for both sides and have discussed which steps are necessary and possible in order to improve the collaboration.

The current presentation supplements these earlier publications in providing a practical hands-on guideline to endangered languages and minority languages corpora for corpus linguists. It gives an overview of existing endangered languages corpora, such as the ones deposited with DOBES, ELAR and PARADISEC, and compares their usefulness and accessibility with regard to corpus research. Relevant properties are e.g., the approximate amount of data per language, the kind of data available, the extent to which these recordings are annotated, the typical kind of annotations used, and the geographical areas they cover. It will be shown, in a hands-on way, how these data can be accessed.
All these criteria are relevant for corpus linguists who want to work with endangered languages or minority languages - in order to determine how useful a specific corpus is for a specific research project, but also as a practical guideline when working with such corpora.
References: