UNIVERSITY OF VERONA, DEPARTMENT “CULTURE E CIVILTÀ”

Formal Representation and Digital Humanities

Text, language and tools

federico.giusfredi@univr.it

29/06/2016

Book of abstracts

Workshop organized within the framework of the Marie Skłodowska Curie Project SLUW “A computer-aided study of the Luwian (morpho-)syntax”; GRANT AGREEMENT NO 655954.

"Reconstructing & Representing": ***New 3D scanning experiments on monumental hieroglyphic inscriptions of Hattusa (Turkey)***

N. Bolatti Guzzo, M. Marazzi, L. Repola, A. Schachner

During the spring of 2014, a cooperative agreement with the mission belonging to the German Archaeological Institute (Abt. Istanbul) at Hattusa, was established, with the aim of experimenting with innovative instruments and procedures in the sector of Cultural Heritage, developed by researchers at CEM (Euromediterranean Research Centre of the University Suor Orsola Benincasa, Naples), within the NOP SINAPSIS ("National Operational Programme for Research and Competitiveness ", NOP 2007-2013).

This agreement had as its primary goal the joint execution, for the first time on the site of Hattusa, of a series of experimentations, which involve essentially:

a. The application of a series of diverse 3D scanner technologies and procedures on monuments of particular artistic and epigraphic value;
b. The development of new procedures for the interpolation and visualization of 3D models in order to identify elements on the monuments the latter not directly visible to the human eye through the observation of the original artefact;
c. The design of new visualization instruments suitable for the scientific research and the epigraphical study of monumental inscriptions;
d. The development of new strategies for the fruition of resulting models inside a museum environment (immersive rooms, augmented reality, etc.).

Special attention was devoted to testing of innovative scanning procedures for constructing 3D-Models of monumental hieroglyphic inscription, whose signs are often no more visible to the naked eye. We present here some preliminary results of the work at 3 different hieroglyphic monuments: the inscriptions on the cliff walls of Nishantash and Yazilikaya (Kammer A) and those on the ashlar wall of Südburg (Kammer 2).

Digital Initiative: The Palaeography of the Anatolian Hieroglyphic Stone Inscriptions

Lorenzo d'Alfonso and Annick Payne

Preliminary studies by both project members have shown the promising scope for detailed research of the palaeography of Anatolian Hieroglyphic Stone Inscriptions. It is of particular importance to achieve reliable dates for the inscriptions which often cannot be dated on either text internal references or archaeological context. As first results show that development strings are not uniform across the corpus of signs of writing but rather show individual developments for each sign, this promises that knowledge of every single development will enable increasingly precise dating with the help of 'dating windows'. Further, a better understanding of developments specific to time and place contains information relevant to the reconstruction of local history, in particular as a means to assess contact with outside groups or isolation. This is of particular importance as the inscriptions under consideration are often the only available text sources for a specific area and period.

The project aims to build up a palaeographic database with photos and drawings which is to be searchable by specific criteria and shall, accordingly, enable the creation of sign lists based one the encoded criteria. At a later stage, research undertaken on the hieroglyphic seal corpus by Dr. Natalia Bolatti Guzzo shall be incorporated in the database. As the project is still in its first phase of conception, the presentation shall focus on project scope as well as digital standards and database solutions which are currently being explored.

Representing Meaning Change in Computational Lexical Resources: the Case of Shame and Embarrassment Terms in Old English

Javier E. Díaz-Vera, Fahad Khan and Monica Monachini

The inclusion of diachronic information that details changes in the meanings of words (or more generally lexemes) over time may be extremely helpful in broad coverage digital lexical resources (and is usually included in more comprehensive general purpose dictionaries), but it is often crucial for lexical resources serving such fields as classical philology or historical linguistics where the diachronic dimension of a language has to be explicitly taken into consideration. The work which we will describe arose from a collaboration between researchers in the field of historical (and cognitive) linguistics and those working on the development of digital language resources and infrastructures; it was the fruit of a shared interest in the formal representation of lexical semantic change, with a view to enabling the development of software tools to analyse and to explore data relating to shifts in word meaning. Our focus is on a specific case study, namely, on the terms used to talk about emotions in Old English (OE) -- with a particular concentration in the initial stages on terms used to talk about shame, guilt and embarrassment -- and on what the changing usages of such terms can us about the conceptual changes were occurring in the English society at the time. The original lexical dataset was an OE lexicon containing information about the various types of meaning shifts in emotion words attested in the OE corpus over a time period stretching from 850 to 1150 AD; our aim was to convert this dataset into more usable format and to facilitate the enrichment of the lexicon with other datasets. We made the decision early on to focus on RDF-based representations of lexico-semantic resources. This choice was motivated by the current popularity of linked data and the semantic web as a means of publishing and linking together datasets, and by the free availability of off the shelf tools for publishing, sharing and querying RDF datasets. We took the *lemon* model for representing lexico-semantic resources in RDF as our starting point. In *lemon* concepts in an ontology are used to stand for the extensions of words and word senses are represented as pairings of lexemes and ontological concepts. Although we made the *lemon* model the basis of our work, it is a general purpose model and so lacks many of the features necessary to represent the salient aspects of the lexicon qua diachronic resource. One difficulty which we had to deal with early on was due to the fact that the syntax of RDF permits the use of only unary and binary predicates -- and in order to represent the change of meaning over time of a word an extra time argument would seem to be necessary.

We will detail our use of the concept of *perdurant* from ontology engineering in order to represent the temporal dimension of meaning shift in RDF (based on our previous work on a diachronic version of *lemon*); more generally we will describe the several different features which we felt it was necessary to add to the *lemon* model in order to better represent the diachronic aspects of the original dataset: both those specific to the OE dataset we were working with, as well as those which characterise diachronic RDF-based lexica in general. We will illustrate our new model by presenting a number of examples of lexical meaning shifts from the OE shame/embarrassment lexicon, along with their representation in RDF using our extension of *lemon*.

On Sonority and Accent in Tocharian B Nominal System

Hannes A. Fellner and Bernhard Koller

Since early 2011 the Linguistics Department at the University of Vienna has hosted a project to create an electronic edition of all available Tocharian manuscript fragments (Comprehensive Edition of Tocharian Manuscripts [CEToM]: univie.ac.at/tocharian). The project aims at offering a unified resource for the study Tocharian texts. It incorporates transcriptions and translations of the texts, as well as detailed commentaries, bibliographical information, links to photographs of the manuscript fragments. In addition we are compiling a comprehensive electronic dictionary of both Tocharian languages.

The present study employs the CEToM corpus to investigate the Tocharian B stress accent in nominal forms. Tocharian B accent can, as is well known, be inferred from the behavior of the central vowel phonemes /ə/ and /a/: the former is rendered by <a> [ʌ] if accented and by <ä> [ɨ] if unaccented, while the latter is rendered by <ā> [a] if accented and by <a> [ʌ] if unaccented. Thus the basic rule for Tocharian B is that disyllabic words bear the accent on the initial syllable, whereas polysyllabic words usually bear the accent on the second syllable (Krause 1952: 10; Krause&Thomas 1960: 43). A number of polysyllabic forms, however, deviate from this general pattern by bearing the accent on the initial syllable. Among these exceptions are a group of forms that according to Malzahn (2010: 6), "have in common that the vowel of the initial syllable is a full vowel such as *ā* or **æ* > TB *e* and [. . .] the vowel of the following, second, syllable is, or was, *(*)ä*"; e.g., ptcp. *eṅku*, pl. *eṅkoṣ* 'seized' < **ćénkəwə*, **-wæṣə*). Recently, Jasanoff (2015) has argued that the synchronic pattern observed by Malzahn constitutes the reflex of a more pervasive phonological Weight-to-Stress Principle operative in a prehistoric stage of the Tocharian languages. According to Jasanoff, Tocharian B stress developed via two stages: "1) replacement of the PIE accentual system by a system of initial stress; 2) advancement of the stress accent one syllable rightwards in words of three or more syllables, except in sequences of the form **-ÁC*$_0$*ə-* (i.e., sequences in which the first syllable contained a "full" (= non-high) vowel and the second contained a schwa or schwa-antecedent (**i*, **u*, **e*, **R̥*))" (p. 90).

Despite being intended as a general account of the development of the Tocharian accent system, Jasanoff primarily focuses on the accentuation of verbal forms. It is the goal of our study to test whether the Tocharian B nominal system provides any evidence for a sonority based account. Specifically, we will determine whether nominal forms with initial accent tend to adhere to the phonological profile laid out by Malzahn and Jasanoff and can therefore be accounted for by a failure to undergo accent advancement due to the Weight-to-Stress Principle. We will use the CEToM dictionary in conjunction with a Perl script to compile a complete inventory of nominal forms whose accent can be determined based on the aforementioned criteria. The resulting dataset will allow us to support or falsify the predictions made by Jasanoff's account in classical Tocharian B, as well as in various other chronological and dialectal layers of this language. Additionally, we will make the resulting dataset publicly available on the CEToM website to provide a basis for future studies on Tocharian B accent.

References

CEToM Comprehensive Edition of Tocharian Manuscripts: univie.ac.at/tocharian

Jasanoff, Jay H. (2015) "The Tocharian B accent", In Malzahn, Melanie et al., eds., *Tocharian Texts in Context. International Conference on Tocharian Manuscripts and Silk Road Culture held June 26–28, 2013 in Vienna*. Hempen: Bremen, 87–98.

Krause, Wolfgang (1952) *Westtocharische Grammatik, Band I. Das Verbum*. Winter: Heidelberg.

Krause, Wolfgang, and Werner Thomas (1960) *Tocharisches Elementarbuch, Band I. Grammatik*. Winter: Heidelberg.

Malzahn, Melanie (2010) *The Tocharian Verbal System*. Brill: Leiden/Boston.

Prince, Alan (1990) "Quantitative consequences of rhythmic organization", In Deaton, Karin et al., eds., *CLS 26-II: Papers from the Parasession on the Syllable in Phonetics and Phonology*. Chicago: Chicago Linguistic Society, 355–398.

(Keynote Paper) The "Digital Philological-Etymological Dictionary of the Minor Ancient Anatolian Corpus Languages" and its contribution to the advancement of digital online encyclopedias

Markus Frank and Zsolt Simon

The great progress which has recently been made in the field of the Digital Humanities recently not only covers the linguistic area of computerized analysis and evaluation of huge text corpora, but also provides far more effective ways for philological work in the field of ancient languages. Thus a digital encyclopedia can be realized as a web interface which holds a wide range of functionality features, coalescing the former distinguished parts dictionary, text corpora and bibliographic databases.

Against this background, the "Digital Philological-Etymological Dictionary of the Minor Ancient Anatolian Corpus Languages" (short eDiAna) attempts to do justice to the new opportunities of the Digital Humanities: we see a philological-etymological dictionary, convenient to operate and adjustable to the individual requirements, as our ultimate objective. Besides its basic functionality as a dictionary, eDiAna will have recourse to an extensive corpus database of the Ancient Anatolian languages together with bibliographical databases containing around three thousand entries (both MySQL database).

The servers of the "IT Group for the Humanities"-Department at Ludwig-Maximilians-University host the whole dictionary (under a CC license). When it comes to information storage and structure, it represents a XML-MySQL hybrid: the running text of each eDiAna chapter is stored as full text XML along with all the tags for the interactive functionality including text formatting. These XML files in turn are filed in a series of MySQL database tables, where they are stored decentrally and in conjunction with the respective meta information. Every time a single dictionary entry is invoked, a PHP construction function selects the relevant elements out of the database, brings them in order, assembles the different parts (integrates bibliographical and / or corpus data if necessary) and delivers the result as HTML output. The technical implementations enable a high performance when operating the dictionary, in addition the output information can be customized effectively to the individual requirements of the actual recipient.

Four different aspects of the eDiAna implementation will be addressed in this lecture:

- A linguistically oriented overview how the dictionaries of the individual Anatolian languages are prepared in this database.
- The public online interface of the dictionary, as far as its range of functions is ready for presentational purposes.
- The input interface that is used to write the single dictionary entries, relying on a heavily modified and enhanced WordPress installation.
- Insights into the fundamental programming mechanics of the dictionary along with tighter information about the developed data structure.

Annotating the syntax of fragmentary texts: the case of Hittite

Guglielmo Inglese

In this paper, I tackle the issue of how to syntactically annotate fragmentary sentences in a Hittite treebank built within the Universal Dependencies' framework (cf. Inglese 2015). Nowadays, though a significant number of Hittite text is currently being digitalized at the *Hethitologie Portal Mainz*, an annotated corpus of the language, let alone a syntactically annotated treebank, is still missing. In Inglese (2015), I set the outline for a Hittite dependency treebank developed within the framework of UD. Crucially, whereas the annotation of linguistic features of Hittite can be easily carried out according to UD's guidelines, with only a few minor adjustments of the template, the encoding of philological notes to texts constitutes quite a challenging task. In particular, cuneiform tablets often attest to fragmentary texts due to physical damage of the manuscripts. Therefore, a schema for annotating fragmentary sentences is badly needed, as otherwise one would be forced to exclude much textual material from the treebank, thereby deeply undermining its representativity. In Inglese (2015), I dealt with the annotation of partially unreadable words, which turns out to be almost unproblematic. Here, I focus on the annotation of fragmentary sentences, i.e. sentences which entirely lack one or more words. The annotation of fragmentary sentences does not constitute a major topic in computational linguistics, and only a handful of scholars have dealt with this issue so far (cf. Zemánek 2007, Korkiakangas & Lassila 2013, and Giusfredi 2015). Building on these works, I discuss how one can exploit UD guidelines to annotate syntactic dependencies within fragmentary sentences in Hittite. In principle, one can either restrict the annotation of syntactic relations to attested tokens only, thus treating broken sentences as a sub-type of elliptical sentences, or one can insert empty nodes to represent missing tokens, providing a syntactic reconstruction of the missing material. As I will show, both approaches present serious flaws, and do not ultimately provide viable solutions for our task. Therefore, I suggest the adoption of an intermediate strategy. I think it is more parsimonious to treat philological gaps within broken sentences as an individual token, and to annotate them accordingly by inserting what I call "structural" nodes. This solution displays several advantages, as it avoids the insertion of empty nodes and speculative syntactic reconstructions, while at the same time allowing for a clear and unambiguous treatment of fragmentary sentences. In this way, one is able to enrich the treebank by including all fragmentary contexts, and yet users can readily keep them distinct from full-sentences, so that any noise in the annotated data is avoided.

References

Giusfredi, Federico. 2015. Phrase Structure and Ancient Anatolian Languages. Methodology and challenges for a Luwian syntactic annotation. In *Proceedings of CLiC-it*.

Hethitologie Portal Mainz, URL: <http://www.hethport.uni-wuerzburg.de/HPM/index. html

Inglese, G. 2015. Towards a Hittite Treebank. Basic Challenges and Methodological Remarks. In *Proceedings of the Workshop on Corpus-Based Research in the Humanities (CRH), 10 December 2015, Warsaw, Poland*, Passarotti, M., Mambrini, F., & Sporleder, C. (eds.), 59-68. <http://crh4.ipipan.waw.pl/proceedings/>

Timo Korkiakangas & Matti Lassila. 2013. *Abbreviations, fragmentary words, formulaic language: treebanking mediaeval charter material*. Proceedings of The Third Workshop on Annotation of Corpora for Research in the Humanities. Sofia.

Universal Dependencies, URL: <https://universaldependencies.github.io/docs/>

Zemánek, Petr. 2007. A Treebank of Ugaritic. Annotating Fragmentary Attested Languages. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories*, De Smedt, K, Hajič, J & Kübler, S. (eds.), 212-218. NEATL: Bergen.

Morphology beyond inflection. Building a wordformation based dictionary for Latin

Eleonora Litta Modignani Picozzi

The computational linguistics world is gradually focussing its interests on researching and building new derivational morphology resources and tools. This happens especially in the production of tools for modern languages such as the lexical network for Czech, DeriNet,[1] and the derivational lexicon for German DErivBASE.[2]

On the Classical languages front, although the number of lexical resources and NLP tools (especially for Latin) is now manifold and varied, until now there has not been any attempt to create a derivational morphology tool, where lemmas are segmented and analysed into their derivational morphological components, so to establish relationships between them on the basis of word formation, and the verbal noun amator can be reconnected to the verb amo through a suffixation of –a-tor. The first steps towards constructing a lexicon based on wordformation for Latin were actually made by Marco Passarotti and Francesco Mambrini in 2012, when they published a paper proposing a model for the semi-automatic extraction of word formation rules and the subsequent pairing of lemmas to their morphologically simplest lemma (i.e. non-derived).[3]

In this context, the Word Formation Latin project (WFL) has been awarded a Marie Curie individual fellowship to expand on these efforts and create a definitive derivational lexicon for Classical Latin. This will ultimately be included in the automatic lemmatiser for Latin LEMLAT (http://www.ilc.cnr.it/lemlat/lemlat/index.html, accessed 21/01/2016, due an update soon), creating a 360° resource for the study of Latin Morphology.

The data is collected and organised in a MySql relational database according to the following steps:

a) A list of lemmas is automatically extracted from the LEMLAT dataset.
b) The wordformation rules (WFR) are conceived according to the Item-and-Arrangement model, which considers word forms either as simple morphemes (simplex) or as a concatenation of morphemes absolving the following conditions:
    1) Baudoin's assumption that both base and affixes are lexical elements (i.e. they are both morphemes),
    2) They are dualistic, having both form and meaning (Bloomfield's "sign-base" morpheme theory)
    3) They both exist in a lexicon (Bloomfield's "lexical morpheme" theory)(PassarottiMambrini, 2012. Hockett, 1954).

In Passarotti & Mambrini, a list of WFRs was obtained both manually and automatically, then identified and formalised into a table, according to their type (prefixal, suffixal, compound and conversion) and according to the category of transformation underwent by the lexical element in input (N-to-N, N-to-V, N-to-A etc.).

In the first phase of the WFL project, for each WFR, we automatically find input and output candidate lemmas through the aid of sql queries (an output lemma can belong to only one WFR).

In phase 2, morphological families are induced from the data. A morphological family is the set of lemmas morphologically derived from one common ancestor-lemma: all those (simple, or complex) lemmas that share the same base are assigned to the same morphological family.

[1] Ševčíková, Magda, and Zdeněk Žabokrtskỳ. 2014. "Word-Formation Network for Czech." In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), 1087–93

[2] Zeller, Britta D., Jan Snajder, and Sebastian Padó. 2013. "DErivBase: Inducing and Evaluating a Derivational Morphology Resource for German", in ACL (1), 1201–11. http://anthology.aclweb.org/P/P13/P13-1118.pdf

[3] M. Passarotti & F. Mambrini, First Steps towards the Semi-automatic Development of a wordformation-based Lexicon of Latin, in Proceedings of LREC 2012, Istanbul, Turkey, 852-859

Finally, the members of each family are automatically linked to each other according to their part of speech, inflectional category, and affixes by means of the WFR assignment. The simple lemma member is assigned the role of ancestor of the family.

This automatic procedure is considered non-ultimate for building the morphological families. However, it provides filtered data that must be checked manually. Manual checking allows the identification of false results, duplication and lacunas resulting from the automatic process; manual hardcoding is necessary for those lemmas produced by poorly productive WFRs, or morphotactically obscure wordformation processes.

For example, in the treatment of the rule that forms nominal adjectives with the addition of the suffix –a-cius/-a-cis/a-x, the sql script pairs and generates two possible candidates for the formation of fugax: fuga and fugium. This duplicate result needs to be analysed and rectified, there must be only one simple input form for each output form, just like there must be only one WFR associated with a derivative lemma.

Evaluation of the language resource is performed by manually checking data organised into homogeneous groups based on WFRs (coverage of rules) and stemming (coverage of morphological families). Precision and recall will be used as evaluation metrics in order to calculate the rate of positive and negative.

So far, 75 prefixal and 97 suffixal rules have been covered, and 17,282 lemmas have been assigned to a WFR.

The quality of precision of the sql queries is higher when the morphotactic mutations are lower: for example in prefixal rules, the precision rate is about 80% to 95%, while in the treatment of the first suffixal rules, precision rate can vary from 75% to as little as 30%. These results are to be considered only temporary, as fine-tuning of queries and a process of exclusion from an ever-growing list of already assigned lemmas can reduce the discrepancy between query-generated results and manual skimming of candidates.

Recall will need to be evaluated at the end of the project, as currently we are unable to verify how many lemmas are not automatically picked up by queries. The presentation will illustrate the methodology employed to obtain the digital resource, the challenges that the Latin language represents as a dead language, the progress through the project schedule and an illustration of mock-up visualisation for the final result.

Subjects, Topics and the Notion of Salience in Indo-European

Rosemarie Lühr

How our database, tagging and multimodal search works will be discussed on the basis of the three terms Subject, Topic, Salience of title. State of the Art In nominative-accusative languages, as the Old Indo-European languages are, the nominative case is marker of the subject, which, according to Keenan (1976), tends to be topic being highly referential. However, Givón (1983) refers to sentences in which there seem to be two clearly distinguishable NPs, one a topic and the other one a subject (John, we saw him yesterday). He then proposes different degrees of topicality, and a separation of subject and topic. Firstly, subject and topic have to be kept apart. Secondly, the subject is traditionally associated with salience, because, cross-linguistically, the creation of relational predicates (encoded in basic transitive verbs) is governed by the following universal principle: the higher argument is more agent-like and more salient in terms of person, animacy and specificity than the lower one.

Cf. Prototypical transitive verbs (Wunderlich 2006).

	λy	λx	VERB(x,y)
abstract case	accusative	nominative	
grammatical function	object	subject	
protoroles	proto-patient	proto-agent	
macroroles	undergoer	actor	
natural distribution of salience (person, animacy, specificity)	less salient	more salient	
natural candidate for	focus	topic	

The interaction of the three notions subjecthood, topichood and salience is not yet sufficiently described for Indo-European. The talk addresses this interplay by analyzing information structure. The central concepts are Topic and Focus, here.

Research questions

Based on a corpus analysis of Hittite, Vedic, Sanskrit, Avestan, Ancient Greek and Latin texts we show (i) under what circumstances the subject and the topic of a sentence coincide (‘topic’ is defined, according to Centering, as the backward-looking center of an utterance; cf. below); (ii) what referring expressions are used to encode subject and topic; (iii) how the notion of salience intervenes.

Database and Tagging

By examining these questions we use the data of our DFG-sponsored projects “Information Structure in Older Indo-European Languages” and “Information Structure in Complex Sentences – Synchronic and Diachronic”. Here, we assume a Topic-Comment and a FocusBackground structure. Contrary to the unitary semantic interpretation of Focus, we presume two kinds of Focus, a New Information Focus and a Contrastive Focus. As for the Topic the theoretical framework is Centering Theory (Grosz, Joshi & Weinstein 1995). Since this theory deals with givenness and salience and as an epiphenomenon with the Aboutness-quality of Topics, both a connection with the Topic-term of the Topic-Comment-structure and above all with the subject is possible. Subsequently, the contextual relations Continue, Retain, Smooth Shift, and Rough Shift will be identified. The Shifting Topic also belongs to the Aboutnessconcept. In the text it changes the perspective towards a new referent. As New-Aboutness Topic it contrasts with the Non-New-Aboutness Topic. In our projects, the different dimensions of information structure are annotated separately. We have developed a basic concept of analysis elements, which allows evaluating the language data with

each other and merging it in a coherent parameter for comparison. All languages are tagged uniformly with EXMARaLDA (http://www.exmaralda.org/en/). Cf. Table (1):

Table (1): Annotated IS parameters

	Tier label	**Content**
1	[text]	word token
2	[lem]	lemma
3	[glos]	glossing, **subject**, **object** etc.
4	[pos]	part of speech
5	**[saliency]**	animacy: human, animate, concrete, abstract etc.
6	**[givenness]**	accessibility: given, new, world-knowledge etc.
7	[definiteness]	definiteness, indefiniteness
8	[context]	identity, anaphora, deictic reference etc.
9	[frame]	scheme according to *Frame Theory*
10	[WPosition]	position for Wackernagel particles, deficient pronouns, auxiliaries
11	[I-particle]	particle which is relevant for information structure, foregrounding particles, backgrounding particles etc.
12	**[shift]**	continue, retain, smooth shift, rough shift
13	**[TOP]**	kind of topic: continuing, shifting, contrastive
14	**[position-T]**	Topic position
15	**[F-domain]**	Focus domain
16	**[NFocus]**	New-information focus
17	**[CFocus]**	Contrastive focus
18	**[position-F]**	Focus position
19	[discourse]	narration, explanation, elaboration, direct speech etc.
20	[style]	stylistic devises, e.g., hyperbaton, tmesis
21	[orig]	sriginal sentence
22	[transl]	German translation
23	[MC/SCclause-st][1]	Main clause status, subordinated elements
23	[MC/SC**grfunct**]	**subject, object, attribute, predicate, adverbials**
25	[MC/SCsyl_no]	syllable number of phrases
26	[MC/SCword-order]	verb first, verb second, verb end, enclitics etc.

[1] MC: main-clause level; SC: sub-clause level/sub-clause-like structure. We are using this term, because we do not only analyze true subordinated sentences.

Such fine-grained analysis units are needed: the unmarked basic structure of a sentence cannot be determined by the focus potential of standard intonation, because in the written corpora of the Old Indo-European languages there are neither phonological, unambiguous morphological, nor syntactical information-structural markers. That means, the prosody of a sentence is only, if at all, indirectly accessible. Also the question test does not help, real questions are much too seldom within the texts. Cf. the example (2):

(2) Chandogya Upanishad 1.10.02 (Lühr 2015)

[text]	sa	ha	ibhyam	kulmāṣān	khādantam	bibhikṣe
[lem]	tad	ha	ibhya-	kulmāṣa-	khād	bhikṣ
[glos]	the: NOM.M.SG	#	rich(M): ACC.SG	cereal(M): ACC.PL	eating: ACC.M.SG	anbetteln:PF.IND.MED3SG
[pos]	prdem	part	noun	noun	prt.prs.act	vfin
[saliency]	pr3.dem/human		human	concrete		
[givenness]	giv		giv	new		
[definiteness]	def		def	indef		
[context]	ana.ref		identity.ana			
[WPosition]		part.XP				
[I-particle]		foreground.p				
[shift]	continue					
[TOP]	Con-T					
[position-T]	initial/pre-part					
[F-domain]			fd			
[NFocus]			nf			nf
[position-F]			post-2P/pre-sub/focus-split			final/post-sub/focus-split
[discourse]	narrator/narrative					
[orig]	sa hebhyam kulmāṣān khādantaṃ bibhikṣe					
[transl]	He asked the rich eating cereals.					
[MCclause-st]	main:decl					
[MCgrfunct]	subj	#	acc-o	pred/acc-o		prsimpl
[MCsyl_no]	1	1	2	6		3
[MCword_order]	#	enclitic	#	#	#	Vend/post-sub
[SC1text]				kulmāṣān	khādantam	
[SC1clause-st]				sub:prt.conj		
[SC1grfunct]				acc-o	v.nominal	
[SC1syl_no]				3		
[SC1word_order]				#	Vend	
[SC1transl]				eating cereals		

The database is ANNIS, an open-source web application that provides access to multi-layer richly annotated corpora and the means for visualizing and retrieving the data [ANNotation of Information Structure for the data of the SFB 632 - "Information Structure: The Linguistic Means for Structuring Utterances, Sentences and Texts", Berlin, Potsdam] (Krause & Zeldes 2014). Pepper is used to import the multiple annotation formats into ANNIS (Zipser & Romary 2010). The data come from syntax, semantics, morphology, prosody, referentiality, lexis and more.

Expected Results

Statistical sampling taken from Hittite, Greek and Vedic demonstrate, that, in fact, there are differences between these languages in combining subjecthood, topichood among each other on the one hand and with different kinds of salience on the other hand. They concern the frequency of the various context relations, the animacy of the subject, the sequence of Topic and subject etc. It also becomes apparent that Hittite goes a special path again (cf. Lühr 2016).

References

Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein (1995): Centering: A framework for modeling the local coherence of discourse. Computational Linguistics 21 (2), 203-225.

Krause, T., & Zeldes, A. (2014). ANNIS3: A new architecture for generic corpus query and visualization. Digital Scholarship in the Humanities.

http://dsh.oxfordjournals.org/cgi/content/abstract/fqu057?ijkey=GJBr0LhNfKW1g8i&keytyp e=ref. Accessed 4 August 2015.

Lühr, Rosemarie (2015): Traces of discourse configurationality in older Indo-European languages? In: Viti, Carlotta (eds.): Perspectives on Historical Syntax (Studies in Language Companian Series 169). Amsterdam, 203–232.

Lühr, Rosemarie (2016): Headedness in Indo-Uralic. In: Alwin Kloekhorst (ed.): The Precursors of Proto-Indo-European' (Leiden Studies in Indo-European). Leiden: Brill / Rodopi [forthcoming].

Wunderlich, Dieter (2006): Argument hierarchy and other factors of argument realization. In: Ina Bornkessel, Matthias Schlesewsky, Bernard Comrie & Angela Friederici (eds.): Semantic role universals: perspectives from linguistic theory, language typology and psycho-/neurolinguistics, 15-52. Berlin: Mouton de Gruyter.

Zipser F., & Romary, L. (2010). A model oriented approach to the mapping of annotation formats using standards. In Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010. Malta.

A digital resource for Greek Linguistics: the Homeric Dependency Lexicon

Silvia Luraghi, Eleonora Sausa and Chiara Zanchi

In this talk, we present the ongoing construction of a new digital resource, *HoDeL* (*Homeric Dependency Lexicon*). *HoDeL* is presently a work in progress. It is being created at University of Pavia (Department of Humanities, Section of Linguistics), and was partly funded by the international project *Argument Structure in Texts*. As the name of the project suggests, *HoDeL* was originally supposed to be a valency lexicon. However, the extracted data can be better described as a dependency lexicon, as we will show in our talk (see also Zanchi, Sausa and Luraghi forthc.).

HoDeL will allow performing different queries on the Homeric poems, starting either from a verbal form or lemma, or from one of its dependents. It will further be possible to refine such queries according to a number of parameters, associated both to verbs (such as lemma, form, voice) and to dependents (such as case, relation, order, type, position, preposition, and conjunction). In its present state, *HoDeL* contains a number of errors, which will be partially illustrated in this talk, and which we are partially reviewing manually.

HoDeL relies on the Homeric Dependency Treebank (HDT), available online at the Perseus Project *Ancient Greek and Latin Dependency Treebank* (AGLDT 1.1, now released in the version 2.0). More specifically, *Hodel* is based on the semi-automatic extraction of all Homeric verbal entries and of their dependents annotated as SBJ (subject), OBJ (object), OCOMP (predicative complement of the object), and PNOM (nominal predicate) at the syntactic layer. The above-mentioned dependents are those said to be part of the verbal valency by the annotation schema of AGLDT (1.1. and 2.0). By contrast, dependents labelled as ADV (adverbials), ATR (attributes modifying nominal heads) and AtV (predicative complements which are not governed by the direct object) have not been extracted.

The data thus obtained have turned out to be problematic in many respects. First of all, *HoDeL* inherited a number of errors due to the fact that the AGLDT 1.1 was manually annotated. Other problematic issues result from the annotation schema adopted by the AGLDT: for example, AGLDT has no dedicated label for null arguments, which are however quite widespread in Ancient Greek (Luraghi 2003, Keydana and Luraghi 2012, Sausa and Zanchi 2015). In addition, the notion of valency as conceived by AGDLT (on the model of the *Prague Dependency Treebank* PDT, see Panevová 1994) hardly seems to be linguistically grounded: for example, it includes passive agents into the verbal valency (against any theory of valency, see e.g. Shibatani 1985, 1988, Dixon and Aikhenvald 2000: 7 sgg., Siewierska 2005, Kulikov et al. 2006: vii-xvii). Finally, the annotation of certain types of dependent is often inconsistent. For example, dative dependents are often annotated as OBJ even when they do not function as second or third arguments of verbs but as adjuncts (i.e. beneficiaries, instruments, locatives), and should be more correctly annotated as ADV.

Websites:

AGLDT 2.0: https://perseusdl.github.io/treebank_data/

AGLDT guidelines: http://nlp.perseus.tufts.edu/syntax/treebank/greekguidelines.pdf

Argument Structure in Texts: https://sites.google.com/site/argumentstructureintexts/home_de

PDT: https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/html/ch02.html

References:

Dixon, R. M. V. and Alexandra Y. Aikhenvald. 2000. *Changing valency. Case studies in transitivity.* Cambridge/New York: CUP.

Keydana, Götz and Silvia Luraghi. 2012. Definite referential null objects in Vedic Sanskrit and Ancient Greek. In *Acta Linguistica Hafniensia: International Journal of Linguistics* 44: 2, 116-128.

Kulikov, Leonid, Andrej Malchukov and Peter de Swart. 2006. *Case, Valency and Transitivity*. Amsterdam/ Philadelphia: John Benjamins.

Luraghi, Silvia. 2003. Definite referential null objects in Ancient Greek. In *Indogermanische Forschungen* 108, 169-196.

Panevová, Jarmila. 1994. Valency Frames and the Meaning of the Sentence. In Philip A. Luelsdorff (ed.), *The Prague School of Structural and Functional Linguistics. A short introduction*. Amsterdam/ Philadelphia: John Benjamins.

Sausa, Eleonora and Chiara Zanchi. 2015. Non accusative null object in the Homeric Dependency Treebank. In Francesco Mambrini, Marco Passarotti and Carolin Sporleder (edd.), *Proceedings of the Workshop on Corpus-Based Research in the Humanities (CHR).* http://crh4.ipipan.waw.pl/proceedings/

Shibatani, Masayoshi. 1985. Passive and related constructions: a prototype analysis. In *Language* 61: 4, 821-848.

Shibatani, Masayoshi (ed.). 1988. *Passive and voice*. Amsterdam/ Philadelphia: John Benjamins.

Siewierska, A. 2005. Passive constructions. In Matthew S. Dryer and Martin Haspelmath (edd.). *WALS Online.* Available at http://wals.info/chapter/107.

Zanchi, Chiara, Eleonora Sausa and Silvia Luraghi (eds.). Forthcoming. *Homeric Dependency Lexicon (HoDeL)*.

(Keynote paper) Well, It Depends. Theoretical and Practical Aspects of the Dependency Turn in Computational Linguistics

Marco Passarotti

After decades of domination by Phrase-based Grammar(s) both in theoretical and in computational linguistics, approaches and applications based on Dependency Grammar(s) have growth substantially over the last 15 years.

The recent publication of 'Universal Dependencies' (a collection of more than 40 treebanks sharing a dependency-based and cross-linguistically consistent annotation style) has confirmed the role of de facto standard played nowadays by Dependency Grammar in the area of language resources and NLP tools.

Such a turn is due both to theoretical and to practical reasons. In my talk, I will deal with both of them.

I will first introduce two of the today most widespread dependency-based annotation styles for treebanks (UDs and PRG), by detailing the theoretical aspects that motivate their different annotation choices.

Then, I will present the results of a number of experiments of mono- and cross-lingual dependency parsing.

Finally, I will show some results of the application of network analysis to manage large dependency treebanks.

The Lexicon of the Neo-Hittite Royal Inscriptions as a Tool for the Analysis of Political Ideology in South-Eastern Anatolian States in the First Millennium BC

Claudia Posani

The Research Project "The Lexicon of the Neo-Hittite Royal Inscriptions as a Tool for the Analysis of Political Ideology in South-Eastern Anatolian States in the First Millennium BC"

was undertaken to achieve a computer-based processing of the Corpus of the Hieroglyphic Luwian Inscriptions Vol. 1, Inscriptions of the Iron Age, Berlin, 2000 by J. D. Hawkins. Its elaboration consists of two main phases:

1. Construction of an electronic database containing the transcriptions and many non-textual

details of all the inscriptions of the Corpus Each word of the inscriptions of the Corpus occupies a cell of a MS Excel©® worksheet together with all the other words belonging to the same paragraph or recognizable text's portion, each of them being placed in a cell of the same Excel©® worksheet line. In this phase it has appeared necessary to adopt standard patterns to homologate transcription's kriteria with those in use in other A.N.E. inscriptions corpora (e.g. the Neo-Assyrian one). Each word, moreover, is listed together with several non-textual data concerning the inscription, such as the place of its discovery, the type and the material of the vectors (e.g., stelae, statues or walls) and many other details, that have been annotated in the same line of the worksheet on which words are set.

2. Creation of an index of the words The elaboration described above enables to obtain, with an internal Office©® procedure, a complete index of all the terms of the inscriptions, with references to the complete sentence to which they belong and to all non-textual data included in the previous phase; at the end of this step the terms will be listed in accordance with all their graphic variants. This step has been completed but is still in need of some adjustments. The complete database enables to realize different kinds of researches that combine textual and support data; e.g, it is possible to carry out researches based on specific lemmata, about which the database provides all the graphic variants and the information concerning the inscriptions to which they belong. In the presentation I will illustrate the phases of elaboration of the database and the way to use it, e.g. to investigate specific rhetorical-ideological aspects.

Parallel training of TreeTagger and RFTagger on Italian CMC linguistic data
Claudio Russo

Nell'ambito di un più articolato processo di trattamento volto alla costituzione di un corpus di italiano nella CMC etichettato per parti del discorso, le sezioni seguenti esporranno i risultati di cinque cicli di etichettatura espletata con i programmi TreeTagger e RFTagger, a fronte di corpora di addestramento progressivamente incrementati, un esteso dizionario di supporto e due tagset definiti in una fase precedente del progetto di ricerca. Da un lato, il presente lavoro intende fornire una comparazione progressiva che sia di orientamento a chiunque intenda intraprendere un percorso autonomo di etichettatura; dall'altro, è parte integrante del percorso strategico del progetto di ricerca in cui è inserito, le cui argomentazioni si rimandano alla sezione conclusiva.

(Keynote Paper) Formal Syntax for Hittite?

Andrej Sideltsev

My contribution will deal with three problems of applying formal models to dead languages, particularly to the syntax of dead languages of the Anatolian group, most notably Hittite.

The first is how to deal with lack of elicitation in a study of a dead language. Most of syntactic phenomena are revealed by batteries of tests for living languages.

The second is how to address lack of negative data.

The third is how to interpret the data, statistically major and minor patterns. Can they indicate *in-situ* and *ex-situ* positions of constituents in view of lack of tests?

The major problem tying in all the problems above is argumentation in favor of a particular solution. Available studies of Hittite syntax within the generative grammar (Garrett 1990; Huggard 2015) very often simply project typological and inner-theoretical data directly onto the syntax of Hittite without giving much inner-Hittite motivation in favor of a solution.

The problems cannot be completely overcome. However, they can be minimalyzed by considering three aspects: (a) statistics; (b) corpus data; and, most importantly, (c) correlating data concerning different constituents. Finally, the results obtained through (a-c) should be evaluated against the attested cross-linguistic variation and inner-theoretical considerations.

I will illustrate the ways the problems of applying formal models to the syntax of Hittite can be overcome by assessing several aspects of Hittite syntax that have been subject to debate in recent time within the minimalism, such as the structural position of preverbs and indefinite pronouns within the Hittite clause, the structure of Hittite vP (NPs/DPs that stay inside vP vs those that raise out of it to Spec, AgrOP, Spec,AgrSP or to Spec,TopP or Spec,FocP, Spec,ForceP).

References

Garrett, A. J. (1990): The Syntax of Anatolian Pronominal Clitics. Ph.D. Diss., Harvard University.

Huggard, M. (2015): *Wh*-words in Hittite. PhD Dissertation, University of California, Los Angeles.

Annotation of Temporal Information on Historical Texts: a Small Corpus for a Big Challenge

Manuela Speranza and Rachele Sprugnoli

Temporal Information Processing (TIP) has the aim of automatically detecting and interpreting events (e.g. actions or situations) and temporal expressions (e.g. dates or periods of time) in texts and identifying temporal relations between them (e.g. an event occurs before another event or a certain date). In recent years, TIP has become an active area of research in the field of Natural Language Processing (NLP), although the resources and automatic systems developed so far cover few domains (e.g. news articles and clinical notes). In the Humanities there is a large research community which could benefit as well from the availability of processing systems for the extraction of temporal information from textual data; in particular, automatic systems could support historical investigation, especially in consideration of the everincreasing availability of digital textual sources.

Since TIP systems often rely on machine learning algorithms which need large amounts of annotated data, many corpora have been released, annotated following TimeML (a markup language created for temporal information annotation [1]). Most of these corpora consist of news articles [2] while for the history domain there is a lack of annotated resources. One exception is the *ModeS TimeBank*, a corpus of Spanish texts dating back to the 18th century manually annotated following the TimeML annotation scheme [3]. In spite of being suitable for use for the development of automatic systems, however, this corpus has so far only been employed for theoretical linguistic studies on the evolution of Spanish.

In view of this gap, we developed the *De Gasperi corpus*, an Italian linguistic resource of 10 articles by Alcide De Gasperi (published in 1914 in the newspaper "Il Trentino") related to the outbreak of World War 1 (see Table 1) [4]. The corpus has been manually annotated by an expert annotator and is freely available for research purposes.[4] Following the TimeML adaptation to Italian [5], the annotation consists 1 of events (e.g. "partire/leave", "rinascita/rebirth"), temporal expressions (TIMEX3s, e.g. "cinquant'anni/fifty years", "1914"), temporal signals (SIGNALs, e.g. "dopo/after") and temporal relations (TLINKs) between two events or between an event and a temporal expression (e.g. the AFTER relation between the events "guerra/war" and "assisteremo/we will witness" in "dopo la guerra assisteremo ad una rigenerazione/after the war we will witness a regeneration").

	Documents	Words	EVENTs	TIMEX3s	SIGNALs	TLINKs
Total Number	10	5,060	1,195	97	62	382

Table 1. The *De Gasperi corpus*: quantitative data

The De Gasperi Corpus has been employed to analyze how well systems built for contemporary Italian perform on historical texts in the context of EVALITA, an NLP evaluation campaign, thus taking the first steps towards a more solid collaboration between the NLP and the Digital Humanities communities with respect to TIP [6].

This initial attempt to develop a complete corpus of historical texts annotated with temporal information led us to highlight two main open issues: 1) can the scarcity of annotated texts be eased by applying crowdsourcing methods as is already happening in other areas of NLP, and 2) is the information conveyed by the TimeML annotation scheme what historians need? As for the first point, the contribution of niche groups of historians rather than the nonexpert crowd could be

[4] http://metashare.fbk.eu/repository/browse/eventitaskevalita2014testgold-pilot/168259104fef11e596150015c5ed7672ca869e 31d8124b5ba05fe96e7835bd4e/

envisaged [7], while a critical analysis of TimeML involving history scholars should be undertaken to deal with the second issue.

References

[1] Pustejovsky, J., Castaño, J.M., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., Katz, G. (2003). "TimeML: Robust specification of event and temporal expressions in text." New directions in question answering. 3: 2834.

[2] Pustejovsky, J., Hanks, P., Saurí, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., Lazo, M. (2003). "The Timebank corpus." Corpus linguistics.

[3] Nieto, Marta Guerrero, Roser Saurí, and Miguel Ángel Bernabé Poveda. (2011). "ModeS TimeBank: A Modern Spanish TimeBank Corpus." Procesamiento del lenguaje natural. 47: 259267.

[4] De Gasperi, Alcide. Scritti e discorsi politici. In E. Tonezzer, M. Bigaran, and M. Guiotto, Vol. 1. Il mulino, 2006.

[5] Caselli, T., Bartalesi Lenzi, V., Sprugnoli, R., Pianta, E., Prodanof, I. (2011) "Annotating events, temporal expressions and relations in Italian: the ItTimeML experience for the Ita-TimeBank." Proceedings of the 5th Linguistic Annotation Workshop. Association for Computational Linguistics.

[6] Tommaso Caselli, Rachele Sprugnoli, Manuela Speranza, and Monica Monachini. (2014) "EVENTI EValuation of Events and Temporal INformation at Evalita 2014". In Proceedings of the Fourth International Workshop EVALITA 2014.

[7] De Boer, V., Hildebrand, M., Aroyo, L., De Leenheer, P., Dijkshoorn, C., Tesfa, B., & Schreiber, G. (2012). Nichesourcing: Harnessing the power of crowds of experts. In Knowledge Engineering and Knowledge Management (pp. 1620). Springer Berlin Heidelberg.

Generating Critical Transcriptions with Character-Based Statistical Machine Translation

Katja Zupan and Tomaž Erjavec

In (digital) scholarly editions, manuscripts are often presented in two transcriptions: the diplomatic one, which tries to faithfully follow the manuscript, and the critical one, which interprets the text, also with the aim of bringing it closer to the modern reader. But while the critical transcription is closer to contemporary language, it often still retains the archaic "flavour"of the diplomatic one, for example with (partial) historical spelling of words. For the authors of critical transcriptions, it could be beneficial to have a computerised method to suggest the correct form of words, thus simplifying and speeding up their work. By now there exists a number of computational methods to modernise (or, in general, normalise) old texts [1], from applying hand-written transformation rules to old forms of words [2], automatically inducing such rules [3], or by using character-based statistical machine translation to translate between archaic and modern words [4, 5]. However, almost all of these methods make two assumptions: they modernise individual lexical units (word forms) and they rely on the existence of a large lexicon of contemporary words in order to filter out potential, but non-existing word hypotheses that the system produces. Both of these assumptions are problematic, esp. in the scenario of critical editions outlined above. First, one of the frequent differences between archaic and modern(-like) spelling is what constitutes an orthographic word, because words that used to be written together are now written apart and vice versa.

Any system that first tokenises the text and then applies transformations to the resulting tokens will fail on such cases. Second, and specifically related to critical editions, a lexicon of contemporary words will most likely not help if the target words are in fact still archaically flavoured.

We propose a simple method, based on statistical machine translation (SMT) [6], to overcome both of these deficiencies. For our experiments we used Moses [6], the de-facto standard open source implementation of SMT. The proposed method is character-based, i.e. it is used to translate characters instead of words. However, in contrast to similar approaches, it does not translate individual words but rather spans of text, in our case individual lines, as these are marked (and aligned) between the diplomatic and the critical transcription. The method relies on a portion of the diplomatic transcription already having the critical transcription, and this pair is then used as the training set to learn the translation model. The target language model is trained on the already available portion of the critical transcription, without recourse to any external data, such as a corpus or lexicon of contemporary language.

We ran a series of experiments to test out this idea, with our dataset derived from the digital edition of "Three sermons on language" by Anton Martin Slomšek [7], a renowned Slovenian bishop. The first of these sermons was written in 1825 and the second in 1829, when Slovenian language had yet to conform to a standardised orthography. Both contain, in addition to the facsimile, the diplomatic and the critical transcription. The first sermon was used for training the CSMT system and the second for testing it. Three sets of methods were explored: the proposed character-level statistical machine translation (CSMT); standard, word-level statistical machine translation (SMT), and the Norma tool [3], which uses a combination of different normalisation techniques. Three variations of language models were tested for machine translation: a) using solely the text of the sermon used for training, b) a combination of a) and the critical transcription of the author's collected works, c) a combination of a) and a modern lexicon of Slovenian language. Evaluation was done by measuring the character error rate (CER), i.e. the average Levenshtein edit distance (difference) between the automatically generated transcription and the manual critical transcription in the test set. As the baseline, we simply took the diplomatic transcription as if it were a critical transcription, which had a CER of 22.62%. The baseline was improved by all methods: SMT reduced the CER to 16.84%, Norma to 14.19% and CSMT to 7.59% in its best setting, on character

bigrams with the simplest of language models (a). The results show that the proposed method, i.e. translating character by character with very basic settings, works best, at least for this particular work: using our method, a transcriber would need to correct about two thirds fewer characters than if starting directly from the diplomatic transcription, making the process of creating a critical transcription more time- and effort-efficient, even when only a small parallel dataset is available for training. On a larger scale, the system could be used to study (non-)systemic linguistic "modernising" interventions to the diplomatic transcription as well as process historical texts with automatic tools for linguistic annotation. In future work, we aim to test the method with further datasets and try outcombinations of various knowledge sources about language in the Moses framework.

References

[1] Piotrowski, Michael (2012). Natural language processing for historical texts (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, vol. 17). Morgan & Claypool Publishers.

[2] Erjavec, Tomaž (2015). The IMP historical Slovene language resources. Language resources and evaluation, 49/3, pp. 753–775, doi: 10.1007/s10579-015-9294-7.

[3] Bollmann, Marcel (2012). (Semi-)Automatic Normalization of Historical Texts using Distance Measures and the Norma tool. In: Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2), pp. 3–14. Lisbon, Portugal.

[4] Tiedemann, Jorg (2009). Character-based PSMT for closely related languages. In: Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT), pp. 12–19. Barcelona, Spain.

[5] Scherrer, Yves, Erjavec, Tomaž (2015). Modernising historical Slovene words. Natural language engineering, doi: 10.1017/S1351324915000236.

[6] Koehn, Philipp (2010). Statistical Machine Translation. Cambridge: Cambridge University Press.

[7] Faganel, Jože; Ogrin, Matija; Erjavec, Tomaž (2004). Anton Martin Slomšek: Tri pridige o jeziku: elektronska znanstvenokritična izdaja. [Anton Martin Slomšek: Three sermons on language: an electronic text critical edition.] Ljubljana: Institute of Slovenian Literature and Literary Studies, ZRC SAZU. <http://nl.ijs.si/e-zrc/slomsek/index-en.html>.