Why linguistics: from a technological perspective

Toni Badia
Universitat Pompeu Fabra, Barcelona

In a technological environment where engines are built that either perform linguistic acts or help users to perform them the question of what linguistics is useful for comes up again and again. And very often the "why" is highly connected to the "what" and to the "how".

The predominant paradigm today in Natural Language Processing (NLP) is data-driven. Since the 1970's, when Artificial Intelligence and NLP were theoretically grounded on logic and linguistic knowledge, a major shift in paradigm has occurred within NLP. The initial success of probabilistic models in speech recognition and machine translation (e.g., at IBM labs: Jelinek at al. 1975; Brown et al. 1990), together with the increase of processing power and storage capacity of computers, has led the field to rely basically on data from which to extract (usually via machine learning techniques) models to perform the tasks at hand. Interestingly this has brought together NLP and corpus linguistics.

The availability of data then is crucial for solving NLP tasks. It has sometimes been argued (Havely et al. 2009) that massive amount of data is enough for success in NLP tasks. Irrespective of the general validity of this statement, the fact remains that for a wide number of tasks there is no massive data available. Models have to be acquired from less data; and the relative scarcity of data can only be compensated if they are richly and carefully annotated, so that satisfying results can be obtained. And indeed in many specialised tasks a combination of massive data with carefully annotated data seems to be required (e.g., when combining in- with out-of-domain data in machine translation).

Why (do we need linguistics)

Linguistics is required to make sense of linguistic facts, which are scattered and intermixed with other communication factors and means. Linguistics has to be grounded on facts, at all levels of granularity: it has to provide explanation of both 1) communication facts taking place in specific time/space coordinates and 2) language as a cognitive phenomenon. Both levels of explanation must be consistent with one another.

- linguistic phenomena are scattered (in multiple linguistic acts and performed in a number of different languages) and intersected with other communication means (gestures, images, intonation...)
- a systematic approach to the correlation between linguistic form and meaning is needed
- linguistic analyses are necessary to provide data that are linguistically structured
- linguistic theories are necessary to provide consistency to the analyses and annotation of data

What (sort of linguistics do we need)

We basically need linguistic theories that are grounded on facts and help in carrying out linguistic annotations of linguistic acts:

- descriptive linguistics must be the core: theory-neutral descriptions, that are applicable cross-language and directly relating form and meaning
- linguistic theories must be validated against (quantitative) data
- linguistics must become a truly experimental science accounting for both: each minute communication act and language as a global communication system (one of the complex systems in nature)

How (should linguistic data be annotated)

Language is a communication means; it is therefore related to any other factor influencing communication (either the channel or the content). Linguistic data have to be treated in parallel to the other elements (image and sound data in audiovisual communication; social data in social media communication...). We need:

- rigorously annotated data (deriving from categorisation of linguistic phenomena)
- using cross-lingual criteria for annotating (so that linguistic diversity is addressed)
- quality annotations (simple, fully justified, grounded on facts, contrasted, replicable)
- clearly quantifiable
- comparable to, and integratable with, non-linguistic data (image, sound, social networks... data)

Brown,P.F. et al. 1990. A statistical approach to machine translation. Computational Linguistics, 16, 2, pp. 79-85

Halevy,A. et al. 2009. The Unreasonable Effectiveness of Data. IEEE Intelligent Systems, Vol. 24, 2, pp. 8-12

Jelinek,F. et al. 1975. Design of a linguistic statistical decoder for the recognition of continuous speech. IEEE Transactions on Information Theory. 06/1975

The Why of an Intelligence Corpus; and How: Ethical and Construction Issues

Robert Buckmaster
University of Latvia

The language of intelligence has become more prominent in everyday discourse, in the mainstream media, on websites and blogs and in the comments sections of newspapers in the first years of the 21st Century as a result of the 9/11 attacks and the subsequent wars and revolutions. This prominence has not been reflected in the literature to any great extent.

Three Cases

On February 4th 2014, a short but explosive four minute ten second tape of an intercepted mobile/cell phone conversation between Assistant Secretary of State Victoria Nuland and US Ambassador to Ukraine Geoffrey Pyatt was uploaded to YouTube (Nuland and Pyatt, 2014) amidst the growing tensions surrounding the ongoing Maidan demonstrations in Kiev, the capital of Ukraine. Most attention was focused on Nuland's colourful 'Fuck the EU' and this phrase was probably why this particular extract of a longer conversation was leaked by persons unknown, though the FSB, successors to the KGB, are suspected. Of more interest to the linguist were usages such as 'deets' for 'details', 'complicated electron', the need to 'glue' things, the necessity for 'an international personality' to 'help to midwife this thing', idiomatic usage like 'we could er land jelly side up on this one', collocations like 'political homework', and mixed metaphors like 'if it does start to gain altitude the Russians will working behind the scenes to try to torpedo it'. The conversation was a fascinating insight into the language of American diplomats at work and a cause of outrage from the offended Europeans.

On the 20th May 2013 Edward Snowden, a former CIA employee and NSA contractor discretely left Hawaii for Hong Kong with four laptop computers and a treasure trove of nearly 2 million documents. He was met in China on the 1st of June by Glenn Greenwald and Ewen MacAskill, Guardian journalists, and Laura Poitras, a documentary film maker. Snowden used a Rubik's cube to make contact with the journalists and was then 'debriefed' or interviewed by the journalists for a week before the first leak was published by the Guardian on the 5th June. Snowden went public on the 6th June and left Hong Kong for Moscow on the 23rd, where his US passport was revoked and he was granted temporary asylum by Vladimir Putin, the Russian president (Wikipedia, 2015). The Guardian and other newspapers continue to print reports based on the Snowden documents. On the 3rd February Greenwald, Poitras and Jeremy Scahill created The Intercept (Greenwald and Scahill, 2015) with a 'short-term mission... to provide a platform and an editorial structure in which to aggressively report on the disclosures provided to us by our source, NSA whistleblower Edward Snowden.'

Looking even further back PFC Bradley Manning leaked a trove of 250,000 US diplomatic cable and 500,000 Army reports to Wikileaks. Manning was arrested in May 2010 and found guilty of violations of the Espionage Act in July 2013 (Wikipedia 2015b.

These three cases illustrate the unprecedented quantity of previously classified material which is now available to the general public. The Manning and Snowden leaks are of an order of magnitude greater than previous leaks like the Pentagon Papers (National Archives), and unlike the Papers are of operational language: the language of internal NSA briefings about classified programmes and the language of State Department diplomats reporting back to Washington in cables.

This paper will address the question of why a corpus of intelligence texts is important and what questions it could answer, as well as discussing the important ethical questions of collecting and analysing leaked/stolen materials - relating this to copyright law, and highlighting some construction issues related to corpus sampling and balance.

References

Greenwald, G. and Scahill, J. 2015. https://firstlook.org/theintercept/about/
Nuland, V. and Pyatt, G. 2014 https://www.youtube.com/watch?v=WV9J6sxCs5k
National Archives http://www.archives.gov/research/pentagon-papers/
Wikipedia. 2015 http://en.wikipedia.org/wiki/Edward_Snowden
Wikipedia. 2015b http://en.wikipedia.org/wiki/Chelsea_Manning

Linguistic Insights for Building Language Technology

Anna Feldman & Jing Peng

Our area of work is Natural Language Processing, which is a field at the intersection of linguistics, computer science, and often many other disciplines. We will describe three ongoing projects that use insights from linguistics.

Morphological analysis, tagging and lemmatization are essential for many NLP applications of both practical and theoretical nature. Modern taggers and analyzers are very accurate. However, the standard way to create them for a particular language requires substantial amount of expertise, time and money. A tagger is usually trained on a large corpus (around 100,000+ words) annotated with correct tags. Morphological analyzers usually rely on large manually created lexicons. As a result, most of the world languages and dialects have no realistic prospect for morphological taggers or analyzers created this way. We have been developing a method for creating morphological taggers and analyzers of fusional languages without the need for large-scale knowledge- and labor-intensive resources for the target language. Instead, we rely on (i) resources available for a related language, (ii) a limited amount of high-impact, low-cost manually created resources, (iii) linguistic observations about the relationship between the source-target morphologies.

The main goal of the second project is to develop a language independent method for automatic idiom recognition. Idiomatic expressions, such as 'a blessing in disguise' and 'kick the bucket' are plentiful in everyday language, though they remain mysterious, as it is not clear exactly how people learn and understand them. There is no single agreed-upon definition of idiom that covers all members of this class, but idioms tend to be relatively fixed in grammatical form and meaning, but with relatively little predictability in the relation between form and meaning. Also, many idiomatic expressions can appear with both literal, i.e. fully predictable, interpretations given their form -- compare 'The little girl made a face at her mother.' (idiomatic) vs. 'The little girl made a face on the snowman using a carrot and two buttons.' (literal). As a result, idioms present great challenges for a variety of NLP applications, including machine translation systems, which often do not detect idiomatic language. To address these challenges, we use linguistic observations combined with machine learning. The starting point is that idioms are semantic outliers that violate cohesive structure, especially in local contexts. The following properties are quantified and incorporated into our algorithm: non-compositionality; violation of local cohesive ties; idiomaticity is not binary -- idioms fall on the continuum from being compositional to being partly unanalyzable to completely non-compositional.

The role of a learner's native language (L1) in second language (L2) acquisition has been widely discussed in the theories of Second Language Acquisition (SLA). The literature suggests that writers' spelling, grammar and lexicon in second languages are often influenced by patterns in their native language. However, the extent of the importance of L1 for acquiring L2 still cannot be determined exactly and remains a controversial topic of SLA research. Recently, the availability of learner corpora has provided opportunities for verifying SLA hypotheses. The previous literature suggests that the best performing features for native language identification are largely the features that rely on the content of the data, such as word n-grams, function words and character n-grams. This means that the future applicability of these features is limited to corpus specific data. The primary goal of our work is to address this problem. We use only non-content based features, part-of-speech tags (POS) and error tags. Exploring these features is useful for corpora independent approaches to native language identification. Our secondary goal is to analyze the features that perform best for highly inflectional data. We approach binary classification as the beginning step in the development of a systematic tool for recognizing a specific L1 from morphologically complex L2 data. We use machine learning techniques to identify features contributing to the classification between Indo-European (IE) and non-Indo-European (NIE) L1 backgrounds of learners of L2 Czech. The results of the experiments show that the non-content based features, especially error tags, are the strongest indicators of the learner's language background.

How corpus linguistics can inform L2 vocabulary instruction: The use of frequency levels

Roger Gee
Holy Family University

Corpus linguistics has contributed to the field of education in numerous ways. The use of word frequency levels in language education has been especially important. Frequency levels are useful in second language (L2) instruction as it is assumed that the more frequent words are most immediately useful, that they are learned first, and that the less frequent words are learned later. Rather than intuition, L2 materials developers and educators have used corpus linguistics to obtain reliable information about frequency levels.

This presentation will report the results of a corpus-based study of the vocabulary used for instruction by Vocabulary.com, itself a corpus-based, freely available game-like site for vocabulary learning (Abrams & Walsh, 2014). Abrams and Walsh suggest that the game-like features of the site promote its afterschool use for vocabulary learning. It logically follows that for words to be learned, the vocabulary used for instruction should not be less frequent than the target words. However, initial inspection of the definitions, usage notes, and sample sentences indicates that these materials may contain a significant percentage of words of a lower frequency than the target word.

It has been argued that for L2 readers, at least 98% of the words (tokens) must be known for adequate comprehension (Nation, 2013). Laufer (2013) reviewed corpus-based research and research involving reading comprehension tests and determined two lexical threshold levels for reading academic, nonfiction texts. One, an optimal threshold level of 8,000 words, would provide about 98% coverage of a text's vocabulary and allow for unassisted reading. Another minimal lexical threshold level of "around 5,000" words (p.809) would not allow for unassisted reading of unsimplified academic texts.

The research reported in this presentation focuses on the 3000-5000 word levels. Words in the 3000-5000 word frequency levels are part of mid-frequency vocabulary (Nation, 2013) and represent a fairly rigorous goal for most language learners. As Webb and Sasao (2013) note, "mastery of the 5000 word level may be challenging for all but advanced learners … the five most frequent levels may represent the greatest range in vocabulary learning for the majority of L2 learners" (p. 266).

A corpus was constructed of the Vocabulary.com defining material for every 20th word of the 3000-5000 word lists from COCA (Davies, 2008-). That is, it begins with word 2001 and ends with word 4981, for a total of 150 words, with 50 from each 1000 word level. The corpus contains the definitions, usage notes, and sample sentences for each of the 150 words. The frequency levels of the defining material for these words were determined using Text Lex Compare and COCA frequency lists.

The presenter will give an introduction to Vocabulary.com, followed by a description of the COCA frequency lists, details of the construction of the corpus, and the method of analysis. To focus on the "why," the results will be contrasted with Laufer's (2013) vocabulary threshold levels. The session will end with time for questions.

References

Abrams, S. a., & Walsh, S. S. (2014). Gamified vocabulary: Online resources and enriched language learning. Journal of Adolescent & Adult Literacy, 58, 49-58. doi:10.1002/jaal.315

Davies, Mark. (2008-) The Corpus of Contemporary American English: 450 million words, 1990-present. Available online at http://corpus.byu.edu/coca/.

Laufer, B. (2013). Lexical thresholds for reading comprehension: What they are and how they can be used for teaching purposes. TESOL Quarterly, 47, 867-872. doi: 10.1002/tesq.140

Nation, I. S. P. (2013). Learning Vocabulary in Another Language. Cambridge, UK: Cambridge University Press.

Webb, S. A. & Sasao, Y. (2013). New directions in vocabulary testing. RELC Journal, 44, 263-277. doi:10.1177/0033688213500582.

Society, Cognition and Language. Understanding human thought and behaviour.

Dylan Glynn
University of Paris VIII

The aim of linguistics is to explain the phenomenon of language – how is it possible that a child learns perfectly a language with seeming ease and possesses an intuitive sense of right and wrong, where philologists have struggled (and thus far failed) for eons to write a grammar that is descriptively and predictively accurate. This is perhaps the only point that all linguists agree upon. Everything from what is language to what is right or wrong is debated and has been debated since the beginning of language science. This talk will argue that, perhaps, this is finally set to change.

Arguably, the fundamental theoretical questions of language science are:
Grammaticality: How do we judge grammatical from agrammatical, and how do we acquire that judgment?
Grammar: What is the structure of language and how to we acquire that structure?

Arguably, the fundamental methodological problems of language science are:
Observability: Neither grammar nor grammaticality are observable. How can we test hypotheses with no direct empirical evidence?
Variability: All the indirect evidence we have reveals significant variation in both grammar and grammaticality. How can we test hypotheses when exceptions are the rule?

The talk will present the usage-based model (Hopper 1987, Langacker 1987) of language and attempt to show how this model may unite the theoretical questions and, in part, overcome the methodological problems.

The usage-based model assumes that there is no grammar of a language *sensu stricto*, but that each individual in a speech community possess a competence. Langauge grammar is, therefore, merely a generalisation across individual competences. In this view of language, the structuring force of language is usage (between individuals of a speech community). In other words, the *langue / competence* is a result of the *parole / performance* and not the reverse, which is assumption sine qua non of structuralist models of language. This reversal of the relationship between structure and use allows us to answer the same fundamental theoretical questions but renders the methodological issues entirely non-problematic.

The talk will present an example of a study that shows how usage-based data can be used to identify conceptual structures relative to their social use. The corpus-driven study will examine the concept of ANGER in American and British English and compare its results with psycholinguistic results. The method, one of many usage-based methods, can be employed to describe any language phenomenon – identifying conceptual structure in behavioural usage. It follows that this method, *mutatis mutandis*, can be employed to test most predictive and explanatory theories of language structure, crossing the *langue-parole* divide.

The Quick Scan: Tailor-made language advice for small and medium-sized enterprises (SMEs)

Hilde Hanegreefs
Zuyd University of Applied Sciences
Mark Pluymaekers
Zuyd University of Applied Sciences

Communication as a bridge between two parties, as a carrier of meaning, is generally -in the business world and beyond- considered to be a means to an end. Successful communication is therefore a powerful tool to influence the knowledge, attitude and behavior of people (The Behavioral Dynamics Institute).

In a time when the information landscape is constantly on the move, high-quality professional communication has clearly acquired a new urgency. Contemporary readers have upscaled their standards: they increasingly require more accessible and transparent texts, both from co-workers within and from business contacts outside of the company. Especially SMEs, where employees often are not sufficiently trained in writing skills, have recently become more aware of the fact that successful communication is critical to building trust, educating and informing stakeholders, influencing public opinion, affecting governmental action, receiving feedback, etc.

The starting point for this project, viz. The Quick Scan, is the growing need and demand for high-quality advice on professional writing, on the part of SMEs. The Quick Scan can be seen as an audit tool that brings to the surface the major problems or shortcomings of a given corpus of written professional communication, using both quantitative (readability test, word count analysis) (De Hertog et al. 2014, Flesch 1951, Scott 1997) and qualitative (discourse analysis, usability tests, spelling and grammar check, channel choice) methods (Renkema 2012). The ultimate goal of this project is to develop a set of scripts describing the methodology with which a particular text type can be ‘scanned’ (e.g., newsletters, press releases, webpages, e-mails and letters). The outcome of the audit will be an advisory report that meets the specific demands of the organization, with the possibility for further follow-up training. In this way, we guarantee tailor-made and research-based language advice in a short time span.

We are currently working on a case study ‘scanning’ the texts aimed at incoming students from five Dutch hotel management schools. The Quick Scan approach will uncover to what extent these texts -as an artifact of a given organization- reflect the schools’ ‘espoused’ or stated values and rules of behavior (Schein 2009) and, as such, help the student in making an informed choice for a particular school. An exploratory quantitative check points out that, in general, the online texts hardly refer to the school’s values and score low on readability. The so-called ‘brochures’, on the other hand, present genre characteristics of folders. Although formally obeying these standards, the visual cues do not always ‘predict’ what is described in the accompanying text. Further usability tests will show whether the content responds to the students’ needs.

The deliverables of this project will not only be beneficiary for business, resulting in more successful corporate communication. It will also provide a better insight into how professional communication ‘happens’ in real-life, into good and bad practices; in short, valuable information that will be taken into account when teaching language skills, such as writing.

References

The Behavioral Dynamics Institute, consulted on January the 26th, 2015: http://www.bdinstitute.org/about-us/

De Hertog, Dirk, Kris Heylen & Dirk Speelman. 2014. “Stable lexical marker analysis: a corpus-based identification of lexical variation”. In: Soares da Silva, Augusto. Pluricentricity: Language variation and sociocognitive dimensions, pp. 127-142. Berlin: de Gruyter.

Flesch, Rudolf Franz. 1951. How to test readability. New York: Harper.

Renkema, Jan. 2012. Schrijfwijzer. Amsterdam: Boom.

Schein, Edgar. 2009. The corporate culture survival guide. San Francisco: Jossey-Bass.

Scott, Mike. 1997. “PC Analysis of Key words”. In: System 25/1, pp. 1-13.

Perception of environmental accountability among representatives of different generations in Russia

Iuliia Goman
St. Petersburg State University

Environmental accountability is gaining more significance in today's world. Its concept is represented differently in environmental research: ecological citizenship (Dobson, Valencia, 2013; Melo-Escrihuela, 2008), environmental justice (Middlemiss, 2010), green citizenship (Smith, 2005), environmentally reasonable citizenship (Hailwood, 2004).

Linguistics can serve as a tool to analyse perceptions of environmental accountability in Russian society. In a wider context this analysis is related to psychology as well as environmental politics. In this research the focus is on the content analysis of programs shown on Russian TV and textbooks used in a business school. This analysis can help to find out the implicit meanings of how a certain issue is shaped and from what angle it is shown for average viewers of TV programs as well as presented in textbooks for would-be managers.

One of the concepts with the help of which environmental accountability is presented is perception of climate change in TV programs. This source of Mass Media is considered well-spread among representatives of Russian society. The perceptions of TV viewers are analysed through the prism of what information a certain TV channel wants them to know. Two TV program s (2010, 2014) on a similar topic were taken as an example to show how perceptions changed, what new concepts were introduced in the debate in a four-year period. Results of content analysis revealed the following : climate change is mostly regarded in connection with catastrophes; it is still not defined as a fact or a public fear; economic benefits from global warming are in focus of discussion.

Another possible source of information for analysis is the content of a textbook for students studying Business English. The aim was to see how the topic of environmental accountability is represented in this source.

The analysis of the second source (the textbook) shows that here the issue of environmental accountability is associated with sustainable development. Company's performance is analyzed from environmental, social, financial dimensions; students learn how to present the position of a group of actors (local government, local business) related to improving the sustainability of the city; they also learn to take real sustainable action (calculating individual carbon footprint).

The conclusion is the following:
1.Being environmentally just to the next generation is a challenge.
2. Information about environmental issues obtained from TV programs is insufficient; further sources of information are needed. Otherwise, an information gap can prevent some layers of society from being well-informed partners in discussions related to environmental accountability, climate change as an example.
3. The discussion of precaution in keeping environment is not typical of an environmental discourse that

takes place via traditional Mass Media sources.

References:

1. Dobson A., Valencia A. 2013. Citizenship, Environment, Economy. New York: Routledge.
2. Hailwood S. 2004. Environmental Citizenship as Reasonable Citizenship. Uppsala: ECPR.
3. Melo-Escrihuela C. 2008. Promoting Ecological Citizenship: Rights, Duties and Political Agency. ACME: Prague.
4. Middlemiss L. 2010. Reframing Individual Responsibility for Sustainable Consumption: Lessons

from Environmental Justice and Ecological Citizenship. Environmental Values 19(2). The White Horse Press.
5. Smith G. 2005. Green citizenship and the social economy. Environmental Politics 14(2). Taylor & Francis.

Understanding Minimalist Communication: A Study of Short Text Genres

Danguolė Kalinauskaitė
Vytautas Magnus University

What?

“Efficient communication relies not on how much can be said, but on how much can be left unsaid” (Brown, Duguid 2000: 205). This is one of the main aspects of minimalist communication (for the term, see Humez et al. 2010). Minimalist communication refers to different forms of communication, that are specific in their greater or lesser constraint on the size. They constitute a vast and miscellaneous group of text genres, short text genres, and even form their own language. The ongoing research is intended to shed light on how particularity of genre influences each separate form of minimalist communication, including both its structure and content (generally speaking). To understand different forms of minimalist communication, their analysis aims to involve the external (a) and internal (b) aspects:
a) form, sphere, and object of communication; communication situation; communication participants; function of text;
b) means of text cohesion; functional sentence perspective; syntactic features; grammatical categories; vocabulary; thematic peculiarities; the way of theme development; etc.

How?

Genres are perceived as “staged, goal-oriented social processes: as social processes because members of a culture interact with each other to achieve them; as goal-oriented because they have evolved to get things done; and as staged because it usually takes more than one step for participants to achieve their goals” (Martin et al. 1994: 233). To make sense of short text genres, they must also be understood as meaningful communicative actions: if we know conventions of particular genres, we are able to do things with language in proper situations and thus our communication is meaningful.

Why?

Some researchers treat text genre linguistics as a separate branch of text linguistics (see Heinemann 2000, Gansel 2011). Given that the notion of genre is “central to understanding the social, functional, and pragmatic dimensions of language use” (Coe, Freedman 1998: 41), the study of short text genres goes beyond the limits of linguistics. It requires interdisciplinarity in both the methodology (a) and theoretical perspectives (b):
a) the study combines the methods and theory of semantics, pragmatics and text linguistics. To collect different forms of minimalist communication in one place and thus represent language use in both the public and private spheres, a corpus of short text genres is created. So corpus linguistics is of specific importance in the study. Next stage, corpus-based analysis of texts, stands on the knowledge of domains of IT and language technologies;
b) the study develops genre theory and thus continues linguistic, literary and rhetorical traditions, as well as ones of language philosophy and communication theory. Naturally, it is also based on them. Due to its touch with communication, the study provides valuable insights not only into humanities, but also into social sciences.

Thus the study of separate forms of minimalist communication contributes to understanding of human communicative behavior in general.

References

Brown, J. S. and P. Duguid. 2000. *The Social Life of Information*. Boston, MA: Harvard Business School Press.
Coe, R. M. and A. Freedman. 1998. Genre theory: Australian and North American approaches.

Theorizing Composition: A Critical Sourcebook of Theory and Scholarship in Contemporary Composition Studies. M. L. Kennedy (ed.). Westport: Greenwood Press. 136–147.

Gansel, Ch. 2011. *Textsortenlinguistik*. Göttingen: Vandenhoeck & Ruprecht.

Heinemann, W. 2000. Textsorten. Zur Diskussion um Basisklassen des Kommunizierens. Rückschau und Ausblick. *Textsorten. Reflexionen und Analysen* 1, 9–29.

Humez, A., N. Humez and R. Flynn. 2010. *Short Cuts. A Guide to Oaths, Ring Tones, Ransom Notes, Famous Last Words, and Other Forms of Minimalist Communication*. Oxford: Oxford University Press.

Martin, J. R., F. Christie and J. Rothery. 1994. Social processes in education. *Language, Literacy and Learning in Educational Practice*. B. Stierer and J. Maybin (eds.). Clevedon: Multilingual Matters. 232–247.

Utilizing Linguistic Resources for Historical Text Clustering

Andres Karjus
Liisi Veski
University of Tartu

Text is the prevalent medium and target of study not only in linguistics, but in a broad range of humanities. When there is plenty of textual data at hand, various computational methods, developed over the last couple of decades, can be used to gain insight into the data, cluster and group the text, model the topics discussed therein, etc. However, when the texts under observation are few and short, state of the art methods appear to perform rather poorly. We propose an idea for supplementing the clustering of small texts by replacing the words they contain using the hyperonymy relations found in a wordnet (a type of lexical database resource, crafted by linguists). The idea of such generaliation, or abstraction, is by no means novel in itself (cf. Hovy, Lin 1999; Durme et al 2009). However, what we propose to use this methodology for is directly reducing the "long tail" of words occurring once or twice in a typycal distribution of words in a text, and cluster the texts using the vectors of the wordnet-generalized terms.

As a case study, we observe the usage contexts of words meaning 'nation, national' in the essays of Estonian scholars and politicians of the 1930's, a time where such matters were hotly debated across Europe. The results show some improvement over the simple bag-of-words TFIDF baseline, but not for all texts. As such, we will discuss possible ways to improve the model.

Hovy, E., Lin, C.-Y., 1999. Automated text summarization in summarist.
Van Durme, B., Michalak, P., Schubert, L. K., 2009. Deriving Generalized Knowledge from Corpora using WordNet Abstraction.

Text Linguistics Approach in Clinical Survey Research and in Diagnostics of Mental Disorders

Svetlana Koudria, St.Petersburg State University, Department of the English Language and Cultural Studies; Russia
Elena Davtian, Russian State Pedagogical University named after A.I. Herzen, Department of Clinical Psychology; St.Petersburg City Psychoneurological Dispensary ¹7, Outpatients' department ¹3; Russia

Problem. The reliability and significance of multinational survey research largely depend on the compatibility of linguistic tools used in different language communities to collect data. One of the major issues in survey research is how to achieve the equivalence of a survey questionnaire translated into different languages. Since the fundamental studies of survey tools lie within the domain of cognitive psychology (2), current clinical questionnaire translation studies describe only individual difficulties in obtaining equivalence, but reveal no systemic causes of those difficulties. Our study seeks to improve translation procedures through the use of the text linguistics approach.

There is also a tendency to use translated clinical research questionnaires for diagnostic purposes in psychiatry. This raises concerns among practicing psychiatrists (1), and makes psychiatrists turn to linguistic theory in search for the rationale and implications of such use of texts. We provide examples of how linguistic theory may be applied in order to support or reject usage of texts as a basis for diagnostics of mental disorders.

Methodology. Our methodology lies within the framework of text linguistics, and suggests two steps: 1. defining a questionnaire as a text type on the basis of the following four parameters: the author and the recipient of the text, the communicative purpose of the text, and the type of information that the text communicates. 2. Providing a functional description of the linguistic constituents of the questionnaire in their relation to the discursive features identified within step one. Using this functional description as a basis for judgments about translation solutions and the diagnostic potential.

Results. The methodology has been applied to 175 clinical questionnaires (4019 questions in total). The clinical research questionnaire has been defined as a highly interactive text with a double pragmatic orientation at two different recipients: the clinician and the patient. The double pragmatic orientation is best illustrated at the level of individual words (e.g. “fatigue”, “frustration”, “anxiety”, “agitation’, “motivated”). The double pragmatic orientation undermines some principles of the standard methodology of questionnaire translation (3), and challenges the use of translated questionnaires for diagnostic purposes.

Conclusions. Text linguistics helps to create a consistent theoretical base for deciding which constituents of the source questionnaire must be retained, and which constituents may be omitted in translation without disrupting the functionality of the translated text. Also, a functional description of the linguistic constituents of the questionnaire reveals the dangers of using questionnaires for diagnostic purposes in psychiatry.

References

1. Davtian E., Koudria S. 2014. A Word in Defense of Clinician (on the use of clinical research questionnaires in psychiatry).//Gannushkin Journal of Psychiatry and Psychopharmacotherapy. Vol. 16, No. 2, P. 59-64.//Moscow: MMA «MediaMedica». (In Russian)

2. Schwarz N., Sudman S. 1996. Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research. Jossey-Bass Publishers.

3. Wild D., Grove A., Martin M., Eremenco S., McElroy S., Verjee-Lorenz A., Erikson P. 2005. Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures. Value in Health. Volume 8. Number 2, P. 94 – 104.

Social unrest through the prism of language: computational linguistic at sociology service

Andrey Kutuzov
National Research University Higher School of Economics
Olga Miryasova
Institute of Sociology, Russian Academy of Science

Deep structure of society is manifested through how people speak or write, and this can help sociologist a lot. However, only recently linguistics and natural language processing developed to the point when they can offer robust methods of analyzing vast amounts of texts to extract meaningful features. We are describing a case when linguistics substantially helped sociology in studying a particular group of grassroots activists.

The group in question is a parents' movement located mainly in Moscow, Russia. In 2012, its participants united around the issue of bad catering in kindergartens. This group consisted mostly of mothers, age 22 to 45, who organized rallies, met with officials and addressed protest letters to the authorities. Activists communicated via their Internet forum at http://forum.materinstvo.ru/, discussing all aspects of their struggle.

This forum became a perfect source of linguistic data about the activists in question. That's why sociologists (initially using participant observation) decided to ask a linguist about what modern natural language processing can do with this data.

From the point of view of a sociologist, analysis of forum posts has the following advantages in comparison to interviews:

1. There are no interviewer's questions, utterances are made because their producers wanted to make them.
2. The utterances are not as artificial as in an interview (activists often tend to feel as if a sociologist' interview is the same thing as an interview to mass media).
3. The amount of texts is much more than in a typical interview (in this case we analyzed a corpus of almost a million words in size, 34 thousand posts, dated from January to December of 2012).
4. It is possible to collect utterances of many people (807 in our case) without substantial efforts.

At the same time, we stumbled upon a few disadvantages:

1. Automatic natural language processing at the level of vocabulary (an even at the level of syntax) often fails to capture cases of ellipsis, irony, extensive usage of euphemisms and co-reference: phenomena which are very frequent in informal collective discussions.
2. Qualitative interpretation of the results of quantitative linguistic analysis is often possible only with the help of an informed expert able to explain apparent inconsistencies in the data and to point at flaws in the text processing pipeline.

In general, sociologists wanted linguistic help with the following issues:

1. To compare relations to power and the authorities within different groups inside the grassroots movement (using textual data as a source). It is important for deciding, whether a particular subgroup perceives their problem as a private one or places it in the wider political context.
2. To refine the exact composition of subgroups picked by an expert. Expert estimation is important, but it is equally important to support it with statistical data about language use by the groups' representatives.

3. To discover key qualitative differences between the subgroups. These differences were found out to be manifested in the distribution of lexical frequencies, as used by the representatives of the subgroups.
4. To estimate how activists' stand changed over time. With traditional methods of sociology, estimation of social movement participants' opinion and behavior dynamics is so cumbersome that it is hardly ever practiced. Massive linguistic analysis of textual data allowed to describe the evolution of people behavior.

All these issues were more or less successfully resolved with the help of natural language processing and machine learning methods. Thus, the collaboration turned out to be fruitful. Computational linguistics allowed to prove sociologists' hypotheses and to come to some unexpected insights.

How Syntax Is Helpful for Statistical Machine Translation?

Huei-Chi Lin
Laboratoire d'Informatique de l'Université du Maine

Phrase-based translation systems are comprised of two probabilistic models, translation model and language model. The translation model is derived from the probabilities of source-target aligned phrase pairs (not linguistic phrase), extracted from parallel corpora. This model is to generate translation alternatives of the given source text, while the language model is taking care of the fluency and grammatical of the translation output. The language model is approximated by the n-gram probabilities trained on huge target language monolingual corpora.

The word alignment is estimated for source-target sentence pair in the parallel corpora, then it is used to extract the consistent phrase pairs for this parallel sentence. The extracted phrase pairs are turned into a phrase table with a larger number of phrase pairs and their probabilities and also reordering table that models short local reordering of words and phrases. Both tables are called the translation model (Koehn, 2010). Translation models have high performance when the two source and target language pairs are close. The translation output therefore only needs short local reorder. The order of syntactic constituents differing between source and target language becomes a critical problem because the parallel data is less monotonically-aligned. The estimation of alignment becomes difficult and translation models are constructed with low quality. For instance, the basic phrase-based systems have less capacity to learn word reordering orientation of English-to-German bitexts. It is not easy to align signal English verbs to German base verbs and their separable prefixes usually locating at the end of the sentence.

One solution for this limitation is to pre-reorder the source sentence to make it resembling the expected order of the target sentence. Hence the translation system needs to do less word movement instead. This is known as "pre-reordering" (Xia and McCord, 2004; Wang et al., 2007; Goto et al., 2012). This preprocessing can be introduced into phrase-based systems with parsing. Syntactic trees have the capacity to represent recursive structures of a language. Reordering rules which are directly learned from parse data can apply maximally on applicable sentences in the source side. When this approach is efficiently performed on source data, the order of translated word respects syntax of the target language.

Syntactic pre-reordering method aims to reconstruct the order of a given source sentence, to make it more close to the order of the target sentence. This syntactic reordering based on pared trees aims to rearrange the maximum pattern in the source data so that the long-distance word reordering is decoupled from the translation systems. The advantage is that the performance of word alignment, reordering models and translation models become better. Translation computed from these statistical systems is hence improved.

References

Isao Goto, Masao Utiyama, and Eiichiro Sumita. 2012. Post-ordering by parsing for japanese-english statistical machine translation. In Proceedings of the50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 311–316, Jeju Island, Korea, July. Association for Computational Linguistics.

Philipp Koehn. 2010. Statistical Machine Translation. Cambridge University Press.

Chao Wang, Michael Collins, and Philipp Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 737–745.

Fei Xia and Michael McCord. 2004. Improving a statistical MT system with automatically learned rewrite patterns. In Proceedings of Coling 2004, pages 508–514, Geneva, Switzerland, Aug 23–Aug 27. COLING.

Expressing uncertainty as a gender-specific use of language: a glance at Estonian data

Liina Lindström
Pärtel Lippus
University of Tartu

The study of gender-specific use of language is a matter of wide interest in different aspects, especially in gender studies and feminist movement, but also in linguistics (see e.g. Lakoff 1973; Tannen 1994). Acknowledging the gender-specific scope of the linguistic variability could hint to different socio-cultural patterns. For example in English it has been shown that female speakers tend to express uncertainty more frequently than male speakers and this can be seen in the frequency of combining first-person singular pronouns with perceptual or cognitive verbs (e.g. Newman et al. 2008).

While in the Estonian society the different position of genders is greater than in most Western European countries (Gender Equality Index, 2013), the gender-specific issues are almost not studied in Estonian linguistics. As a consequence of the unbalanced gender status in Estonian society we expect the insecurity that the female Estonians experience in their every-day life to be reflected in their verbal communication. But rather than diving deep into fine-grain discourse analysis, we aim to demonstrate the tendencies to sociolinguistic patterns with very basic frequency-based observations. Similarly to the findings from English (Newman et al. 2008), we expect male speakers to make statements as they were given facts and female speakers to stress their personal opinion.

In this paper we take a look at five different modal expressions that are used as hedges in English (Newman et al. 2008). These expressions add uncertainty to the statement thereby soften or reduce its force. Firstly, we look at three different variations of modal adverbials marking self’s opinion: *minu meelest*, *minu arust*, ‘in my oppinion’, and *minu jaoks* ‘for me’. Secondly, *ma ei tea* ’I don’t know’ expresses not knowledgening, however, it is used often as a modal particle. The frequent use of it may indicate that in the speech situation the speaker is lacking of self-confidence, and thus female speakers could also use it more frequently. Finally, the confirmation asking particles *eksju* and *onju* ’isn’t’, which are somewhat similar to English tag-questions that also are thought to be more idiomatic for female speakers (Lakoff 1973), as asking for confirmation is another expression of uncertainty.
The data comes from the University of Tartu Phonetic Corpus of Spontaneous Speech (http://www.keel.ut.ee/foneetikakorpus). The dataset comprises transcriptions of spoken dialogues recorded from 69 speakers (34 male and 35 female; about 30 minutes each; about 200 000 words in total). The occurrences of the five expressions under study were analyzed as a function of speaker’s gender and age, and the interlocutor’s gender and age.

The results show different patterns for all observed expressions. The constructions *minu arust* and *minu jaoks* are used by younger speakers regardless of their gender, while *minu meelest* is used more by female speakers. The use of expression *ma ei tea* revealed a more complex pattern: there is no general gender difference, but both male and female speakers use it more when speaking to a partner of their own age and gender. The particle *onju* is more frequently used by young female speakers, while *eksju* in the corpus seems to be idiomatically frequently used by a small number of speakers.

References

Gender Equality Index – Country Profiles, 2013. http://eige.europa.eu/content/document/gender-equality-index-country-profiles
Lakoff, R., 1973. Language and woman’s place. Language in Society, 2(01), p.45.
Newman, M.L. et al., 2008. Gender Differences in Language Use: An Analysis of 14,000 Text Samples. Discourse Processes, 45(3), pp.211–236.

Tannen, D., 1994. Gender and discourse, New York: Oxford University Press.

Why put 'bio' in bio-linguistics?

Pedro Tiago Martins
Pompeu Fabra University
Cedric Boeckx
University of Barcelona/ICREA
Constantina Theofanopoulou
University of Barcelona
Javier Ramirez
University of Girona
Elizabeth Zhang
University of Barcelona
Gonzalo Castillo
University of Barcelona
Edward Shi
University of Barcelona
Saleh Alamri
University of Barcelona
Anna Martinez Alvarez
University of Barcelona
Gokhan Dogru
University of Barcelona
Evelina Leivada
University of Barcelona

Advancements in theoretical linguistics, especially since the mid-20th century, under the impetus of generativism (Chomsky 1957, et seq.), have led to the important insight that human language is a biological capacity, often called the language faculty. The way this insight has guided research, however, has not yielded a plausible biological characterization of this faculty. Much theoretical machinery has been devised by linguists sympathetic to the idea of a language faculty, and linguistics has indeed become a much richer field in the process, but there has been little effort to come to grips with genuine biological concerns. One can find mentions of biology, genetics, and evolution in some of the theoretical linguistic literature, but it seems that they are confined to introductory sections, and that most of the work amounts to language description, albeit in a sophisticated fashion. The relation between linguistics and biological disciplines has remained, at best, metaphorical. Lenneberg (1964: 76) was correct in saying that "[n]othing is gained by labeling the propensity of language as biological unless we can use this insight for new research directions-unless more specific correlates can be uncovered."

The inadequacy of linguistics in offering a biological account of the language faculty is far from being widely acknowledged—quite the opposite—leading to proposals that are not tenable from outside of linguistics. The very limited degree of success in finding brain correlates to linguistic primitives is a good example of the disparity between the kind of work that has been routinely carried out in linguistics on the one hand, and neuroscience, on the other. Poeppel & Embick (2005) illustrate this state of affairs by pointing out two problems: i) *granularity mismatch*: "Linguistic and neuroscientific studies of language operate with objects of different granularity. In particular, linguistic computation involves a number of fine-grained distinctions and explicit computational operations. Neuroscientific approaches to language operate in terms of broader conceptual distinctions", and ii) *ontological incommensurability*: "The units of linguistic computation and the units of neurological computation are incommensurable". These problems arise from the insular fashion in which lingustics has operated, with great aims but little interdisciplinary dialogue.

We argue that it is only by fully embracing the biological sciences that these problems can be surpassed, and that linking hypotheses can be put forward towards the goal of a plausible biological characterization of language. Our illustrations will draw from several disciplines, such as evolutionary biology, cognitive neuroscience, and comparative psychology, and will lead to a reconsideration of the sort of data linguistics must wrestle with if they are serious about putting 'bio-' in their business cards.

References

Chomsky, Noam. 1957. Syntactic Structures. The Hague: Mouton.
Lenneberg, Eric. 1967. A Biological Perspective of Language. In Eric Lenneberg (Ed.), New Directions in the Study of Language. Cambridge, MA: MIT Press.
Poeppel, David.; Embick, David. 2005. Defining the relation between linguistics and neuroscience. In Ann Cutler (Ed.), Twenty-first century psycholinguistics: Four Cornerstones. Mahwah, NJ: Lawrence Erlbaum.

From Coenoses to Style and Harmony: on Interdisciplinary Potential of Linguistics

Gregory Martynenko
St. Petersburg State University

1. The paper deals with the interdisciplinary potential of linguistics, revealing its relations to the general theory of coenoses, to the mathematics of harmony, and to a diverse number of metrical disciplines. The author intentionally skips well-known and evident interrelations of linguistics with other neighboring disciplines and focuses on less-researched interdisciplinary relations. The paper is unique in a way that the author uses his own research experience and includes only those interdisciplinary interactions that the author faced himself during his long and versatile research career path. That gives an opportunity to look at linguistics from new different perspectives and expand its potential further.

2. General theory of coenoses is an interdisciplinary theory that deals with community and population studies. This theory has an origin in biocoenosis (or ecosystems) studies that describe interacting of organisms living together in a common habitat. The theory was expanded and enriched by principles used in systems theory, classification theory, sociological theory, community theory, and theory of statistical population. You may also find coenoses or communities in different scientific fields (e.g., biogeocoenoses, technocoenoses, urban coenoses, linguistic coenoses, etc.).

Any text (both written or oral) being an integral unity consisting of multiple elements, not necessarily homogeneous, may be considered as a specific kind of coenosis. Text integrity is achieved by cohesiveness of author's ideas and unity of theme, plot, style, and other essential factors. The set of words, sentences or paragraphs, which constitute text, can be regarded as its lexical, syntactic and hyper-syntactic "population" [1].

Regardless of the particular subject area, the researchers use common systemic notions and terms for description of communities, e.g., such terms as homogeneity-heterogeneity, stability-instability, balance-disbalance, order-randomness, concentration-dispersion, integrity- amorphousness, complexity-simplicity, etc. Interdisciplinary studies of coenoses use common mathematical models and methods of data processing within one conceptual framework.

3. Stylometrics (or stylometry) is a philological discipline associated with studies of linguistic style. It has its origin in the works of German philologist W. Dittenberger dedicated to the problem of anonymous text attribution. Stylometrics ideas have much in common with metrical studies in other scientific areas: biometrics (F. Galton and K. Pearson), psychometrics (G. Fechner), "art-metrics" (A. Zeising), biometrics, anthropometry methods used to identify criminals ("bertillonage" by A. Bertillon), econometrics (V. Pareto), and others. At the end of the XXth century the goals of stylometrics were reformulated expanding its area to the broader set of tasks associated with ordering and systematization of texts and their components in regard of stylistic features (e.g., taxonomy, attribution, dating, morphology, periodization, diagnosis, identification) [2].

4. Mathematics of harmony is an interdisciplinary research area that is based on the synthesis of the theory of harmonic proportions and the theory of recurring sequences (such as Fibonacci sequences). In linguistics, it may be used for studying text composition, poetic structure, word frequency lists, rhythmic structures, etc. Mathematical concept of recursion was introduced in linguistics in form of syntactic ideas and the theory of generative grammar by N. Chomsky. When studying syntactic structures, this concept was expanded by adding specific linguistics content. Thereafter, this extended notion of linguistic recursion was brought back to mathematics where it is now used for typology of Fibonacci sequences. Thus, we observe the interdisciplinary migration and evolution of the term "recursion". The interdisciplinary aspect of this study is reinforced by the fact that these mathematical

and linguistic structures can be introduced to any other scientific discipline where Fibonacci numbers are used (economics, medicine, architecture, music, etc.) [3].

References

1. Martynenko, Gregory. 2009. Chislovaja garmonija lingvocenozov [Numerical Harmony in Linguistic Coenoses]. In: Martynenko, Gregory. Vvedenie v teoriju chislovoj garmonii teksta [Introduction to the Theory of Numerical Harmony of Texts]. St. Petersburg: St. Petersburg State University.
2. Martynenko, Gregory. 1988. Osnovy stilemetrii [The Foundations of Stylometrics]. Leningrad: Leningrad State University.
3. Grigoriev, Yuriy, and Martynenko, Gregory. 2012. Tipologija posledovatel'nostej Fibonacci: teorija i prilozhenija. Vvedenie v matematiku garmonii [Typology of Fibonacci's Sequencies: Theory and Applications. An Introduction to the Mathematics of Harmony]. LAP LAMBERT Academic Publishing Gmbh & Co. KG.

Linguistics in semiotics: metaphorical models and methodological innovations

Katre Pärn
University of Tartu

Linguistic theories have had explicit and noteworthy impact on semiotics since the inception of the 'semiology' by Ferdinand de Saussure, who on one hand defined linguistics as a branch of the more general science of semiology - since language is only one particular semiological system among others; on the other hand stated that linguistics can become the master-pattern - or model - for all branches of semiology. To a great extent the semiological project did became the application of linguistics to other sign systems. Although sometimes this "linguistic project" is seen as a historical approach that bears little relevance in contemporary semiotics, underneath the varying terminologies of new developments in semiotics we still often find linguistic influences and theories.

For Saussure, the reason why linguistics should have had this role in semiology/semiotics had to do with the arbitrary nature of linguistic sign that made language a special kind of social institution. When some 50 years later Roland Barthes redefined the relationship between linguistics and semiotics, envisioning the latter becoming a part of trans-linguistics instead, it was mainly because of the centrality of linguistic mediation in culture that does not allow any sign system to bypass the relay of language.

However in closer inspection, the answer to why linguistics has played such an important role in semiotics cannot be reduced to the special status of language or nature of linguistic sign, but has more to do with "how" of the linguistics - how linguistics as science models language as its object of study and constructs linguistic theories as descriptive and explanatory paradigms. This specific, perhaps in the context of humanities even somewhat revolutionary epistemological attitude towards its object of study and way of theory-building is also rooted in Saussure’s Course that outlined this new attitude and provided the theoretical model of language to be used in the study of other sign systems as language-like systems.

I will discuss the key aspects of this model-based approach to point out why linguistics has been in past and still is today through its new developments an important source of methodological innovation for semiotics.

The Role of Linguistics in Improving Statistical Machine Translation as Scientific Field and as End-Result

Tommi A Pirinen
Antonio Toral
Dublin City University

In statistical machine translation (SMT), the main goal behind most of the research work to date is to improve translation quality as measured by automatic evaluation metrics (e.g. BLEU, METEOR, TER). These metrics are cheap as they are fully automatic and they are useful in that they (are supposed to) correlate with human translation evaluation when measured on large bodies of professionally translated texts. That said, a mere race for tiny improvements in these metrics (i.e. there is a common type of paper in the field that (i) adds some novel functionality to the overall SMT pipeline and (ii) reports a statistically significant improvement in terms of BLEU as evidence of its usefulness, without additional human evaluation nor in-depth analysis) has not been overly successful with all of the SMT language pairs.

This seems to be especially the case for some of the non Indo-European languages, e.g. Finnish or Turkish. An example of this is reflected in the state-of-the-art of machine translation for the Finnish–English language pair, which has not advanced in the past decade in terms of these metrics. These metrics are based on simple string comparisons and substitutions, and as it is argued in the papers published on this language pair that show little to no improvement[1, 2, 3], these metrics may not be ideal for morphologically complex languages like those of the Uralic or Turkic families. For example, a simple compounding mistake or a wrong allomorph for case suffix is severely penalised by BLEU (basically, a word matches or does not match the reference) while the fluency and meaning in the translation is mostly retained.

We believe that using metrics that include linguistically-motivated measures (e.g. MEANT [4]) would be more appropriate, especially when translating into the morphologically complex language. In addition, we propose to use metrics that combine matching at word and sub-word levels. In this regard, taking the most widely-used automatic metric, BLEU, we suggest to combine its original implementation (i.e. word level) with its proposed implementation at morpheme level (m-BLEU[2]).

Apart from the issue of non-linguistic automatic evaluation metrics, we deem equally crucial to analyse the output of the SMT system in sufficient detail to explain the scores achieved by the system. From a meta-analysis of the state-of-the-art in English–Finnish SMT that we have conducted, we identify two parts of the process that would greatly benefit from introducing a linguistic view, namely (i) the initial hypothesis when devising a new SMT system and (ii) the final error analysis of the MT output. In fact, in most papers (i) the experimental set-up is not motivated by any relevant linguistic means and (ii) error analysis is not conducted at all, let alone in linguistic detail. These, we argue, would be necessary to design future work aimed at improving current results in terms of real translation quality.

References

[1] Ann Clifton and Anoop Sarkar. Combining morpheme-based machine translation with post-processing morpheme prediction. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11, pages 32--42, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[2] Minh-Thang Luong, Preslav Nakov, and Min-Yen Kan. A hybrid morpheme-word representation for machine translation of morphologically rich languages. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP '10, pages 148--157, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[3] Sami Virpioja, Jaakko J. Väyrynen, Mathias Creutz, and Markus Sadeniemi. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In Proceedings of the Machine Translation Summit XI, pages 491--498, Copenhagen, Denmark, September 2007.

[4] Chi-kiu Lo and Dekai Wu. Meant: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility via semantic frames. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pages 220--229. Association for Computational Linguistics, 2011.

Putting Text-Level Linguistics into Statistical Machine Translation

Andrei Popescu-Belis
Idiap Research Institute; Swiss Federal Institute of Technology in Lausanne

Statistical machine translation systems are quite successful at translating individual sentences, in particular when sufficient training data is available for a given language pair. However, these systems do not yet take advantage of the relationships between the sentences of a text, and hence do not yet ensure a coherent translation of entire texts. In this talk, I will show how to make available text-level linguistic knowledge to a phrase-based statistical machine translation system. This approach, which has been pioneered by a Swiss-based consortium of linguists collaborating with language engineers, will be exemplified on three types of phenomena: discourse connectives, verb tenses, and noun phrases. I will explain how theoretical and data-driven linguistic modeling has guided the design of automatic labeling modules, which enabled machine translation systems to generate more coherent translations of entire texts.

The Austrian Baroque Corpus ABaC:us: What does the linguistic annotation add?

First Author: Claudia Resch
Other Author: Eva Wohlfarter
Austrian Centre for Digital Humanities, Austrian Academy of Sciences

The term "corpus" in linguistics refers to a large and structured set of texts which is usually electronically stored and processed. The purpose of this paper is to introduce the Austrian Baroque Corpus (ABaC:us) which has been built up by an interdisciplinary team since 2010.

ABaC:us consists of text data and images dating from the baroque era, in particular the years from 1650 to 1750. It includes 17 texts with more than 210.000 running words, of which five texts - attributed to the Augustinian monk Abraham a Sancta Clara (1644-1709) - constitute the very core of the corpus. The texts of ABaC:us belong mainly to the so-called Memento Mori genre, thus to texts associated with death and dying.

The corpus aims to combine traditional philological expertise and up-to-date text technology to preserve the cultural and linguistic heritage embedded in the texts. In order to ensure reusability, well-established text technological standards - XML annotations according to the guidelines of the Text Encoding Initiative (version P5, http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf) - were adopted. The focus of the paper, however, lies on the linguistic annotation: With Tree Tagger (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/), an open standard to apply Part of Speech tagging, and the Stuttgart-Tübingen-Tagset (http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf) word class and lemma information were automatically added to every word in the five main texts of the corpus. But what does the linguistic annotation add to the value of the corpus? The question is legitimate, as the manual correction of the annotation - which was necessary to obtain high quality data - was a rather time-consuming process.

The linguistic annotation allows for more complex linguistic research, such as the analysis of stylistic and rhetorical features, recurring patterns and grammatical elements. Can the linguistic analysis of the corpus help us to enable a deeper knowledge of the society of the past? With several examples from ABaC:us, this paper aims to open the debate.

References

Boot, Peter. 2009. Mesotext. Digitised Emblems, Modelled Annotations and Humanities Scholarship. Amsterdam: Pallas Publications, Amsterdam University Press

Czeitschner, Ulrike, Declerck, Thierry, and Resch, Claudia. 2014. Porting Elements of the Austrian Baroque Corpus onto the Linguistic Linked Open Data Format. In: Osenova, Petya, Simov, Kiril, Georgiev, Georgi and Nakov, Preslav (eds.): Proceedings of the Joint Workshop on NLP & LOD and SWAIE: Semantic Web, Linked Open Data and Information Extraction associated with the 9th International Conference on Recent Advances in Natural Language Processing (RANLP 2013). Sofia: p. 12-16

Dipper, Stefanie. 2010. POS-Tagging of Historical Language Data: First Experiments. In: Semantic Approaches in Natural Language Processing. Proceedings of the 10th Conference on Natural Language Processing (KONVENS-10). Saarbrücken: p. 117-121

Hinrichs, Erhard, Zastrow, Thomas. 2012. Linguistic Annotations for a Diachronic Corpus of German. In: Linguistic Issues in Language Technology, Volume 7, issue 7, p. 1-16

Kawaguchi, Yuji, Minegishi, Makoto, Viereck Wolfgang (eds.). 2011. Corpus-based Analysis and

Diachronic Linguistics. Amsterdam and Philadelphia: John Benjamins

Moerth, Karlheinz, Resch, Claudia, Declerck, Thierry and Czeitschner, Ulrike. 2012. Linguistic and Semantic Annotation in Religious Memento Mori Literature. In: Atwell, Eric, Brierley, Claire and Sawalha, Majdi (eds.): Proceedings of the LREC 2012 Workshop: Language Resources and Evaluation for Religious Texts. Paris: ELRA, p. 49-52

Resch, Claudia, Declerck, Thierry, Krautgartner, Barbara and Czeitschner, Ulrike. 2014. ABaC:us revisited - Extracting and Linking Lexical Data from a historical Corpus of Sacred Literature. In: Atwell, Eric, Brierley, Claire and Sawalha, Majdi (eds.): Proceedings of the 2nd Workshop on Language Resources and Evaluation for Religious Texts / LREC 2014. Reykjavik: p. 36-41

Perfectivity As An Attention Phenomenon: The Aspectual Theory Expects Neural Evidence

Nezrin Samedova
Azerbaijan University of Languages

1.1. The idea that aspectual meaning has a specific cognitive nature already appears in a paper published in 1891 [1: 379]. In 1927 work, S. Karcevski writes that perfectivization is nothing else but the concentration of our attention on one concrete moment of a process that excludes all other moments and creates the impression that the perfective process has no duration at all [2: 89]. Cf. [3].

Having imbibed classical ideas, the theory I follow states that perfectivity is the seme ‘punctual bound’. The seme has a peculiar cognitive nature. Attention focuses on it and, consequently, the seme backgrounds some semantic features of a verb, the seme ‘process’ in particular. Very importantly, the ability of backgrounding has a fixed power. The perspective corroborates and explains, among others, the scholars’ intuitions about the following cases in Russian verbal system.

2.1. Verbs like prygnut’ (“jump”) are traditionally treated as punctual / punctive, instantaneous, momentary, i.e. as having no seme ‘process’. Indeed, cf. prygnut’ k stolu and doprygat’ do stola. On the other hand, one cannot deny that such verbs possess the seme ‘process’, cf., e.g. [4: 185, 227] and phrases like медленно прыгнул (“jumped slowly”).

The paradox is solved due to differentiating the homonymous imperfectives prygat’1 and prygat’2 (“jump”). We show that they significantly differ regarding their behavior. The latter refers to a very short physical action that does not last long. The action is conceptualized in full accordance with the characteristic and Russian captures the concept, namely prygat’2 possesses the peculiar seme ‘process having short (i.e. non-standard) duration’. This is also true for prygnut’, while both prygat’1 and doprygat’ have the seme ‘process of standard duration’.

Thus, the illusion that prygnut’ is punctual emerges because it is the seme ‘process of short duration’ that is overshadowed by the aspectual seme ‘final bound’, whereas in doprygat’ the seme ‘process of standard duration’ is backgrounded. Cf. the visual metaphors:

doprygat’ do stola prygnut’ k stolu

————————————● ————●

2.2. When comparing units like stat’ prygat’ and zaprygat’, some scholars note that the construction foregrounds the process, cf. e.g. [5: 221]. We show that stat’+INF, unlike the verb, is characterized with the syncretic seme ‘process of long (i.e. non-standard) duration’ and the seme ‘initial bound’ is not able to conceal it completely.

2.3. The seme ‘initial-final bound’ makes interesting effects in perfectives like poprygat’ and proprygat’.

3. Thus, the theory provides linguistic evidence regarding one case of the interaction of semantic elements, namely the existence of attention phenomena in language, cf. [6]. We expect neurolinguistic studies to reveal the underlying neural mechanisms of the described phenomena.

References

1. Размусен Л.П. (1891). О глагольных временах и об отношении их к видам в русском, немецком и французском языках // Журнал министерства народного просвещения, С.-Петербург, № 6, с. 376-417.

2. Карцевский С.И. (2004). Система русского глагола / С.И. Карцевский. Из лингвистического наследия. Т. 2, Москва, с. 31-207.

3. Петрухина Е.В. (2000). Аспектуальные категории глагола в русском языке в сопоставлении с чешским, словацким, польским и болгарским языками. Москва.

4. Маслов Ю.С. (1959). Глагольный вид в современном болгарском литературном языке (значение и употребление) / Вопросы грамматики болгарского литературного языка. Москва, с. 157-312.

5. Зализняк А.А., Шмелёв А.Д. (2002). Семантика ‘начала’ с аспектологической точки зрения / Логический анализ языка. Семантика начала и конца. Москва, с. 211-224.
6. Talmy, L. (2007). Attention phenomena. In: Handbook of Cognitive Linguistics. Ed. By Geeraerts, D. and Cuyckens, H. OUP. P. 264-293.

Studies of Everyday Speech at the Intersection of Disciplines

Tatiana Sherstinova
St. Petersburg State University

The research project described in the paper has been started several years ago with the aim to investigate Russian spontaneous speech. As it was shown by many researchers, natural speech is very different from speech recorded in a laboratory with soundproof walls. We decided to change the conception and get recordings from natural real-life communicative situations: the participants-volunteers had to spend a whole day with turned-on voice recorders that recorded all their audible communications. This methodology can be compared with a daily cardio monitoring that is widely practiced in medicine. As known, speech features strongly depend on speaker's individual characteristics. Thus, we incorporated into our analysis techniques used in field linguistics, sociolinguistics, and psycholinguistics (e.g., the participants had to complete socio-demographic and psychological questionnaires). That was the origin of the linguistic resource later known as the ORD corpus of Russian everyday communication [1].

The ORD corpus allows to examine speech on various linguistic levels: phonetic, lexical, grammatical, semantic, and pragmatic. Moreover, the detailed examination of corpus data led us to unexpected conclusions: the ORD recordings give valuable research data for many other interdisciplinary studies like anthropological linguistics, behavioral and communication studies, studies in pragmatics, discourse analysis, psycholinguistics, and forensic phonetics. The corpus can also be used for didactic purposes when studying colloquial Russian as a foreign language, etc. This paper focuses on the two important ORD applications.

Applications for Speech Technologies. Statistical description of Russian spontaneous speech in everyday interaction is very significant for adjustment and improvement of speech synthesis and recognition systems. Thus, speech transcripts of the ORD corpus may be used for building n-gram language models for speech recognition systems that predict the probability of a given word on the basis of the preceding n−1 words. The study of spontaneous speech reduction [2] may be used for building an authentic lexicon of word pronunciations. Besides, the specialists in speech technologies express interest in the lists of the most frequently used Russian utterances [3] and in temporal patterns of speech obtained from the ORD data.

Sociolinguistic studies. Several pilot sociolinguistic investigations were made in a last few years based on the ORD data: e.g., speech rate studies, comparison of men’s and women’s social behavior, etc. Recently, a large sociolinguistic project has been started with an aim to analyze everyday Russian with focus on social differentiation (age-, gender-, education-, professional-related groups, etc.). Sociolects are to be described on phonetic, lexical, and grammar levels. One of the most significant objects of this project is to reveal distinctive speech features between different social groupings (e.g., young people vs. older people, men vs. women, blue collars vs. white collars, etc.). Besides sociolinguistics, the results of this project will be a very valuable material for forensic linguistics as well. The research is supported by the Russian Scientific Foundation, project # 14-18-02070 "Everyday Russian Language in Different Social Groups".

The list of possible applications of the ORD corpus may be further continued.

References

1. Asinovsky, Alexander et al. 2009. The ORD Speech Corpus of Russian Everyday Communication "One Speaker's Day": Creation Principles and Annotation". Text, Speech and Dialogue, LNCS/LNAI

5729. vol. 5729, 250–257. Berlin/Heidelberg: Springer-Verlag.

2. Bogdanova, Natalia, and Palshina, Daria. 2010. Reducirovannye formy russkoj rechi (opyt leksikograficheskogo opisanija). Slovo. Slovar'. Slovesnost': Tekst slovaria i kontekst leksikografii. Mat. Vseross. nauchn.Konferencii, 491–497. St. Petersburg: RGPU imeni A. Gerzena.

3. Sherstinova, Tatiana. 2013. Russian Everyday Utterances: the Top Lists and Some Statistics. Dialogue Studies. Vol. 20. Approaches to Slavic Interaction, 105–16. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Why (evolutionary) linguistics?

Mónica Tamariz
University of Edinburgh

Linguists study the structure of human languages. Evolutionary linguists see this structure as the result of individuals' cognitive biases unfolding over social interactions. We are interested in how these genetically specified cognitive biases evolve biologically, how the properties of languages evolve culturally and how these two evolutionary dynamics interact. Evolutionary linguistics, straddling biological and cultural processes, can thus help us understand who we humans are: Proximally, as social individuals, we are the product of the niche we ourselves create: culture, including language. Ultimately, as a species, we are the product of co-evolutionary interactions between language (a very important part of the environment where our genes have evolved) and our genes (which have evolved to make the most of that environment by making us extraordinary learners and users of languages). Within this framework, this talk will focus on experimental evidence of the cultural-evolutionary processes that explain linguistic structure. I will conclude by arguing that a full explanation of the properties of languages must take into account cultural processes like transmission and communicative interaction.

Named Entity Recognition in Estonian

Alexander Tkachenko
University of Tartu

Knowledge confined within natural language can be made more accessible for machine processing by means of transforming the text into a structured, normalised database form. Information Extraction aims to do just this - its goal is to automatically extract structured information from unstructured text documents using natural language processing. One basic sub-task in Information Extraction involves the recognition of predefined information units such as names of persons, organisations, locations, and numeric expressions including time, date, money and percent expressions. Named Entity Recognition (NER) is the process of identifying these entities in text.

In this work, we discuss common issues related to building a NER system using supervised learning framework. Specifically, we aim to investigate an effect of language-agnostic and language-specific features on system performance. In NER, language-agnostic features are largely based on character makeup of words and include information such as prefixes, suffixes, capitalisation, etc. Language-specific features, however, are based on words' grammatical and morphological information. These include, for instance, word's lemma, part of speech, case, etc. Although language-specific feature have been shown to result in a higher performance, they require availability of sophisticated tools such as morphological analyser or part of speech tagger, which are not available for many less popular languages.

Additionally, we present our recent findings in using unlabelled text to boost NER performance.

As a result of experimentation, we achieved an overall F1-score of 87%, which is compatible with results reported for similar languages.

Dynamics of ethnolinguistic vitality of Maris

Elena Vedernikova
Institute of Estonian and General Linguistics, University of Tartu

Current research is about ethnolinguistic vitality of Mari, a Finno-Ugric minority of Russia (total number is around 550, 000). Mari is a slowly decreasing ethnic group (around 3% every 25 year), mainly due to linguistic and ethnic assimilation (Ehala&Vedernikova 2015). The concept of ethnolinguistic vitality was defined as "that which makes a group likely to behave as a distinctive and active collective entity in intergroup situations” (Giles et al 1977) and involved such disciplines as social psychology, sociology and linguistics. For last 30 years the theory of vitality was refined significantly, having extended significantly its interdiscipliniarity.

There are numerous research works concerning various aspects of historical, linguistic, and cultural development of Maris. Based on the descriptive knowledge, some authors make attempts to assess the strength of Mari people in the Russian society but they lack empirical basis for which the the question "How vital is Mari people in Russian society?" had been left unanswered. Meanwhile, if to take into account the social and political situation in Mari El (autonomous republic), this issue is vital for Maris themselves.

The research is based on the ethnolinguistic vitality theory of Martin Ehala (2010), whose mathematical model allows to calculate the vitality and answer the question "What is the vitality of Maris?" The main components of the given model are: 1) the perceived strength differential PSD; 2) perceived intergroup discordance (D); 3) perceived intergroup distance (R); and 4) the level of utilitarianism (U). In collecting the data, survey questionnaire of ethnolinguistic vitality (Ehala 2009) was used. It included 60 questions using Likert scales. The results of the quantitative study were deepened by qualitative analysis (interviews) of public discourse in Mari El with focus group interviews.

The whole data was collected during three fieldworks in 2013-2014 (Mari El, Russia), and processed by SPSS statistical package (14.0 version).

As according to Austin and Sallback (2010), 50-90% of current distinct languages will become extinct by the 2100, then the issue of preservation of multilingualism and multiculturalism of the world becomes more essential. There are numerous language revitalization projects all over the world, and one of the goals of scientific community is to assist their activities by exploration of vitality of an ethnic group in order to elaborate a high-efficient revitalization program. The case of Mari can be a good example for conducting a similar research of other minorities both in Russia and outside of it.

References

Austin, Peter K; Sallabank, Julia. 2011. Introduction. In Cambridge Handbook of Endangered Languages. Cambridge University Press.

Ehala, Martin. 2009. An evaluation matrix for ethno-linguistic vitality. In S. Pertot, T. Priestly and C. Williams (eds.), Rights, Promotion and Integration Issues for Minority Languages in Europe. Palgrave Macmillan.

Ehala, Martin. 2010. Refining the notion of ethnolinguistic vitality. International journal of multilingualism.

Ehala, M. Vedernikova, E. 2015. Subjective vitality and patterns of acculturation: four cases. Journal of Multilingual and Multicultural Development.

Giles, Howard, Bourhis, Richard, Taylor, Donald. 1977. Towards a theory language in ethnic group relations. – Language, ethnicity and intergroup relations. London: Academic Press.

Statistical methods for particle verb extraction from text corpus

Eleri Aedmaa
University of Tartu

Series of studies have been conducted on using association measures (AMs) to identify lexical association between pairs of words that potentially form a holistic unit, but the question "what is the best AM?" is still difficult to answer. It was unknown how the AMs perform on Estonian data and which AMs are most successful for collocation extraction. This study focused on a subtype of collocations or multi-word expressions, namely particle verbs – a frequent and regular phenomenon in Estonian and problematic subject in natural language processing. I tried to ascertain the best AM for the extraction of particle verbs through investigation of the impact of corpus size on the performance of the symmetrical association measures and compared symmetrical association measures t-test, mutual information, X2, log-likelihood function and minimum sensitivity and asymmetrical conditional probability and ΔP. t-test achieved best precision values, but as the corpus size increased, the performances of X2 and minimum sensitivity improved. In addition, I demonstrated that ΔP is successful for the task of particle verb extraction and provides us slightly different and more detailed information about the extracted particle verbs.

Teaching Business Communication Skills in English: Analysis of the Content

Goman Iuliia
St. Petersburg State University

Since application of the method of content analysis is wide and multifunctional, it seems possible to apply it to teaching business communication skills in English. The analysis comprises materials on the topic 'Environmental Sustainability'; it will help to demonstrate the fact that while acquiring business communication peculiarities in English not only do students improve language competence, but also they acquire a higher level of analytical and critical skills as well as more advanced awareness of environmental issues that they may encounter as representatives of society. This research aims at analysing texts and tasks' formulation and finding out proofs of advancing students' analytical and critical skills as well as their awareness of environmental issues. Improving analytical and critical skills is possible by means of formulating relevant tasks that students as future managers have to fulfil; organizing self-study (additional reading); presenting cases and finding solutions to them. Improving environmental awareness is the result of a proper upbringing that is achieved while studying the course in question and thanks to selecting the right materials for reading; fostering 'think climate', 'think green' approaches to environment; taking a responsible and active position in life. The results of the research will show a set of ways of improving analytical and critical skills (read about catastrophes, environmental issues; propose solutions to solve cases) as well as environmental awareness (learning how to count carbon footprint, selecting materials to read for making a presentation). Implication of content analysis of the materials for teaching will provide an insight into better organization of the course that improves a range of a future manager's skills.

Predicting constructional choice in Estonian

Jane Klavan, Maarja-Liisa Pilvik, Kristel Uiboaed
University of Tartu

A common presumption in usage-based linguistics is that the alternation between linguistic rival forms (such as the English genitive constructions) is not free but conditioned by a multitude of factors. In our presentation, we take a closer look at two near-synonymous constructions - the synthetic adessive construction (e.g. laual 'on the table') and the analytic "peal" construction (laua peal 'on the table') - expressing spatial locative function in Estonian, and identify a number of semantic and morpho-syntactic factors that influence the choice between the two constructions.
In the first systematic study on the subject (Klavan 2012) a logistic regression model with four morphosyntactic and two semantic explanatory predictors was fit to Estonian written language data, yielding a classification accuracy of 70%. In our study, we use dialectal data from the Corpus of Estonian Dialects (CED 2015) to explore how the minimal adequate model for written data performs on non-standard, spoken spontaneous language. In addition, we include the geographical dimension and the Landmark lemma as random effects and demonstrate how these factors significantly improve the fit of the model. Furthermore, we show how complementing the results of the mixed-effects logistic regression model with the results obtained with the 'tree & forest' models (e.g. Breiman 2001) helps to explain the variation in more detail and highlight significant interactions in the data.

References:
Breiman, Leo. 2001. Random Forests. Machine Learning 45(1): 5-32.
CED. 2015. Corpus of Estonian Dialects. http://www.murre.ut.ee/mkweb
Klavan, Jane. 2012. Evidence in Linguistics: Corpus-Linguistic and Experimental Methods for Studying Grammatical Synonymy (Dissertationes Linguisticae Universitatis Tartuensis). Tartu: University of Tartu Press.

Hedging in the peer review process

Roger Michael Alan Yallop
University of Tartu

Peer review is becoming widely accepted as an effective and commonly used teaching method on Academic L2 writing courses (Leijen and Leontjeva 2012). However, participants are often uncomfortable at directly criticising their colleague's written texts in their review comments. So as not to offend their peers, they often hedge their reviews in order to mitigate or soften their criticisms. Crompton (1997) in his summary of the literature explains that hedging is a common linguistic devise used as a politeness strategy to make 'things fuzzier' as a threat minimizing strategy. This presentation explores how one dyad use hedging devices to mitigate their review comments on one another as part of an on-going longitudinal study. I explain how I use Salager-Mayer's (1994) taxonomy to code the pair's hedging devices. I present the results graphically and speculate on how this knowledge can improve the effectiveness of the peer review process.

Bibliography

Crompton, P. (1997). Hedging in academic writing: Some theoretical problems. English for Specific Purposes, 16/4: 271-287.

Leijen, D. and Leontjeva, A. (2012). Linguistic and review features of peer feedback and their effect on implementation of changes in academic writing: A corpus based investigation. Journal of Writing Research, 4/2: 177 - 202.

Salager-Meyer, F. (1994). Hedges and textual communicative function in medical English written discourse. English for specific purposes, 13/2, 149-170.

Corpus Studies of Russian Everyday Speech and Oral Communication

Bogdanova-Beglarian Natalia, Sherstinova Tatiana, Blinova Olga, Martynenko Gregory
St. Petersburg State University

The paper presents the ORD ("One day of speech") corpus of Russian everyday speech which contains long-term audio recordings of daily communication [1]. Nowadays, the ORD corpus is the most representative collection of everyday spoken Russian containing more than 1000 hours of recordings gathered from 110 main participants and hundreds of their interlocutors; speech transcripts numbers about 500000 words and it is planned to extend transcripts up to 1 million words. Speech is selectively annotated on different levels — phonetic, lexical, grammatical, and pragmatic; quantitative data processing is made for annotations on each level [2]. The paper gives brief overview of studies which are (or have been) conducted on the ORD data in the followings aspects: 1) phonetics (study of reduction; temporal studies; speech patterns; hesitations; etc.); 2) lexical studies (new words; new meanings; frequency word lists; lexical richness and concentration; slang; argot; etc.); 3) morphology studies (POS-distribution; frequency lists of grammatical forms; grammatical errors; etc.) 4) syntactic studies (linear word order; syntactic complexity; specific syntactic phenomena of spontaneous speech; etc.); 5) discourse and communication studies (macro and micro structures of everyday communication; communication scenarios; discourse words and fillers; pragmatic studies; communication with "not-standard" interlocutors; etc.); 6) psycholinguistic studies (dependency of speech characteristics from speaker's psychological type); and 7) sociolinguistic studies (speech features of different social grouping; gender linguistics; styles and registers of spoken Russian; etc.) currently supported by Russian Scientific Foundation, project # 14-18-02070 “Everyday Russian Language in Different Social Groups” (cf., for example, [3]). The ORD corpus has different interdisciplinary applications, the major of which will be listed.

References

1. Asinovsky A., Bogdanova N., Rusakova M., Ryko A., Stepanova S., and Sherstinova T. (2009) The ORD Speech Corpus of Russian Everyday Communication "One Speaker's Day": Creation Principles and Annotation". Text, Speech and Dialogue, LNCS/LNAI 5729. vol. 5729, 250–257. Berlin/Heidelberg: Springer-Verlag.
2. Sherstinova T. (2010) Quantitative Data Processing in the ORD Speech Corpus of Russian Everyday Communication, In: Grzybek, P., Kelih, E., and Mačutek, J. (eds.) Text and Language: Structures, Functions, Interrelations, 195–206. Wien: Praesens Verlag.
3. Bogdanova-Beglarian N., Asinovsky A., Blinova O., Markasova E., Ryko A., and Sherstinova T. (2014) Zvukovoj korpus russkogo jazyka: novaja metodologija analiza ustnoj rechi [Sound Corpus of Russian: New Methodology of Oral Speech Analysis]. In: Jazyk i metod: Russkij jazyk v lingvisticheskikh issledovanijakh XXI veka. [Language and Methodogoly. The Russian Language in Linguistics Studies of the XXI-th Century]. Krakow: Jagiellonian University (in print).

General Language as an Acquired and Organized set of Phrases

Sudharsan R Iyengar, Theron Rabe
Winona State University, Winona, MN, USA

Abstract:
A person’s language (vocal as well as written) that is used in communication and in thinking and reasoning is the resultant set of acquired, practiced and organized phrases over their lifetime. This is similar to one utilizing acronyms and phrases in ones field of expertise. Music and chants are similar but may not require semantics to go along with it. A formulation of these phrases together with associated semantics is used for communication. Hence, the myriad of languages and their “grammar”. We are working on using lambda calculus as a universal language processor. We have implemented a programming language EESK (available on Github) and working on enabling it to assimilate a “self-evolving” general language and consequently perform inductive, deductive, and abductive reasoning.

References:
Turing, A. M. Computability and λ-Definability. The Journal of Symbolic Logic. Vol. 2, No. 4 (Dec., 1937), pp. 153-163.
Reynolds, J. C. (1993). The discoveries of continuations. Lisp and symbolic computation, 6(3-4), 233-247.
Frontiers in Artificial Intelligence and Applications, Vol. 171, AGI 2008, pp. 409-413.

How do African Americans speak in Finnish?

Tomi Paakkinen
University of Turku, Finland

My doctoral thesis examines the translation of African American English into Finnish in literature. The research material consists of seven translations of seven novels by seven different African American authors. The purpose is to analyse the translations for the statistical frequencies of occurrence of lexical, morphological, phonological and syntactic features of colloquial Finnish. In a previous study, Sampo Nevalainen (2004), using a large corpus, discovered that in translations, lexical features of speech (i.e. colloquial words and expressions) were most commonly used to create the illusion of spoken language, whereas in fiction originally written in Finnish, phonological features of speech were most commonly used for that purpose. In my master's thesis, I studied the translation of African American English into Finnish in three original novels and their translations and found that lexical features of spoken language were the most prominent in only one of the three translations, whereas in the other two translations, phonological features of speech were the most prominent. In the two novels that produced an unexpected result, the black characters differ from the white characters in terms of their personality, and for these characters, African American English is a strong marker of identity. The aim of the study is not only to discover how African American English is translated into Finnish, but also to discover whether characterisation in original novels has an effect on the translators' use of language.

References

Nevalainen, Sampo. 2004. Colloquialisms in translated text. Double illusion? Across Languages and Cultures 5, 67–88.