Parallel historical corpora – a new method in standardization research?

Dr Marija Lazar Saxon Academy of Sciences lazar@saw-leipzig.de

The Saxon and Magdeburg Law as a cultural link between the legal orders in the Eastern and Middle Europe

Project Head: Prof Dr Heiner Lück www.magdeburger-recht.eu

Ivs maideburgense in oriente: preliminary considerations

The number of parallel historical texts is very restricted and therefore sets the limits for comparative studies in historical linguistics. The researcher of *Mirror of the Saxons* (*Sachsenspiegel*) and *Magdeburg Law* (called together *ivs maideburgense*) is at an advantageous position. Being an important law source, *ivs maideburgense* was adopted in many Middle and Eastern European countries in the Middle Ages. It was translated into the languages used there and therefore influenced their legal systems and languages (Lieberwirth 1986).

In many cultures, the legal register is typologically one of the first in which the writing is established and elaborated (Reutter 1982). Having a wide public reach, it crucially impacts the development of the whole language by its nature and therefore deserves the most attention. The key notion of this development seems to be the *standardization*. The Slavic translations of *ivs maideburgense* are used to explore this process against the background of the language contact.

Žilina Law Book (*Žilinská pravná kniha,* 1378-1561; ed. by Kuchar 2009 and Piirainen 1972) is one of the most important sources of *ivs maideburgense* in the Western Slovak area. It contains a copy of *Mirror of the Saxons* in Middle High German (further MHG) as well as its translation into 'slovakized' Czech (sCz) and provides comparative material for the study of standardization. Because the empirical evidence for the Czechoslovak linguistic continuum is still not satisfactory (Berger 1997), *Práwa saszká* (1469-1470) from the Northern Bohemia is used to compare the vernacular variation in these two sources of *ivs maideburgense*.

Standardization

Standardization is a matter of interdisciplinary interest between the history of law and linguistics.

From the linguists' point of view two notions of standardization have to be distinguished (cf. Kopaczyk 2012; my own emphasis – *ML*):

1) a stage in the development of a language designated through "prestige, formal stability and functional versatility" (*language* standardization);

2) a quality of a language when "a specific repertoire of *acceptable*, or even *expectable*, constructions and phrases" is established (*linguistic* standardization).

For the law historians standardization is a quality of a language used in legal discourse, which ensures its *comprehensibility* for the participants of this specialized discourse.

The comprehensibility is achieved through usage of the *conventionalized* expressions in certain genres and textual patterns (Kjær 1991).

Obviously, the *linguistic* standardization is the cutting point between the history of law and linguistics and the study of the *conventionalized multi-word units* is relevant for both disciplines. The following description of the legal language will focus on them.

The fact that *ivs maideburgense* is a translated law adds the further layer to the analysis design. The law is expressed in a particular language, which makes up its cultural identity and as a result processing of its concepts within the network of this particular law (Kjær 1995). Thus the reception of a new law is inevitably an impetus for language contact and two language systems have to be taken into consideration. This challenge is met by creating *parallel corpora*.

The parallel historical corpus

The parallel historical corpus is applied as a method for linguistic standardization research. It provides a suitable environment for identification and comparison of the legal textual chunks in the large context (paragraph) in both languages.

The corpus based on *Žilina Law Book* unifies the approaches to *parallel* and *historical* corpora. The workflow encompasses nine steps as shown in the following diagram and varies across the languages involved according to the input:

German	'slovakized' Czech	Format	Programmes/scripts/macros
digitizing of the existing paper edition		.pdf	
language recognition		readable .pdf	Adobe Acrobat Pro
OCR	OCR	.rtf or .docx	Adobe Acrobat Pro
OCR-output correction	OCR-output correction	.rtf or .docx	TUSTEP
word-by-word-comparison with the paper edition		.rtf or .docx	TUSTEP/Word
standardization	standardization	.rtf or .docx, .txt	TUSTEP
alignment of the paragraphs	alignment of the paragraphs	.txt, .pws	ParaConc
metatextual tagging	metatextual tagging	TEI	TUSTEP
conversion	conversion	.xml	NoSke

For *Práwa saszká* the concordance in Word has already been made for the previous studies upon the legal terminology within our project. Standardization, alignment, metatextual tagging, and conversion for this concordance are necessary.

n-grams: an indicator of standardization

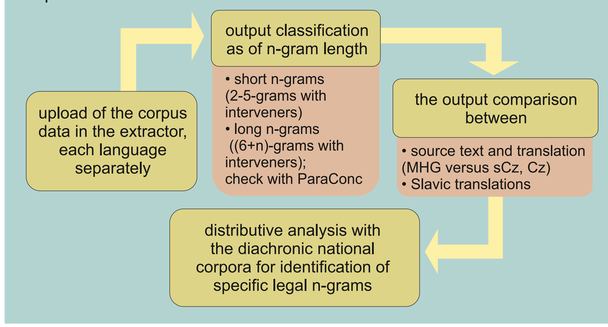
Idea

The n-grams are textual chunks/multi-word units, which are frequently used in certain linguistic domains, genres, and even textual patterns and are thus symptomatic for *linguistic standardization* process within them (Kopaczyk 2012). Because our knowledge of the legal domain in the Czechoslovak linguistic continuum is limited, we are first interested in the whole stock of the n-grams we can gain from *Žilina Law Book* and *Práwa saszká*. Then, the corpus-based distributive analysis will provide the insight into the core area of the legal language in both sources.

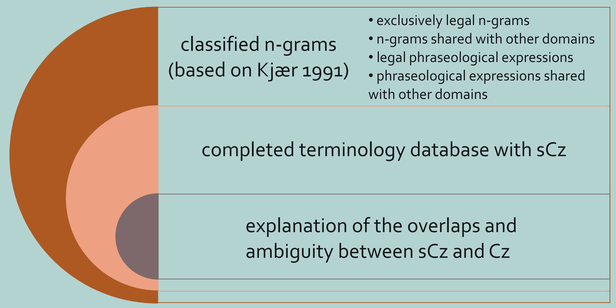
Tools

- N-Gram Phrase Extractor (http://lextutor.ca/n_gram/)
- ParaConc
- www.korpus.sk
- www.korpus.cz
- Vocabularies and lexical databases

Implementation



Output



Case study

Harvesting

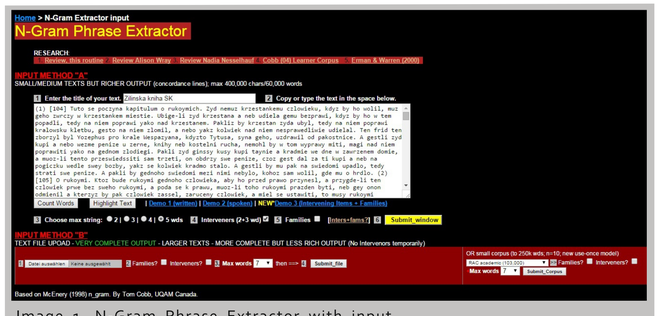


Image 1. N-Gram Phrase Extractor with input
Input: *Žilina Law Book* test corpus (approx. 2700 tokens)
Search parameters: 2-5-grams with interveners

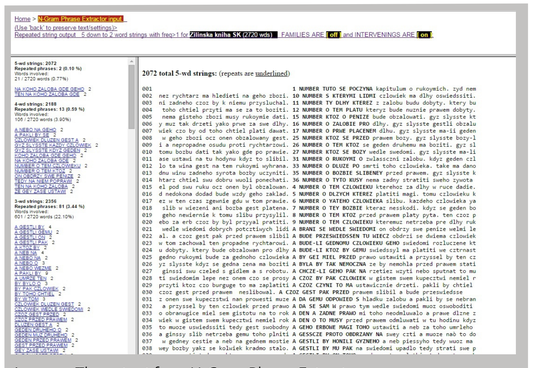


Image 2. The output from N-Gram Phrase Extractor

Apples and oranges: sorting out

Long legal n-grams

sCz:	MHG:
ten [INT], na koho zaloba gde, geho erbownik/erbowe	Stirbet aber iener auf den dy clage get sein erbe
'the one, who is charged, his heir/heirs'	'If the one dies, who is charged,his heir/heirs'

Legal phrase

sCz:	MHG:
ten/onen/czlowiek, na koho zaloba gde	iener auf den dy clage get
the one, who is charged' (lit.: 'the one, upon whom the claim comes')	'the one, who is charged' (lit.: 'the one, upon whom the claim comes')

Short n-grams shared with other domains

sCz:	Cz:
A gestli by / A pakli by 'if'	Synthetic verbal forms with inversion – bude li; neda li; nema li
MHG: Inversion	'if the one is / does not give / does not have'

Boiling down

- MHG contains more n-grams

n-grams	Middle High German	'slovakized' Czech
2-grams	494 (15,89%)	317 (12,11%)
3-grams	228 (7,81%)	81 (3,44%)
4-grams	100 (3,52%)	13 (0,59%)
5-grams	68 (2,4%)	2 (0,10%)

Graph 1. The n-grams in the test corpus based on *Žilina Law Book* (absolute and weighted)

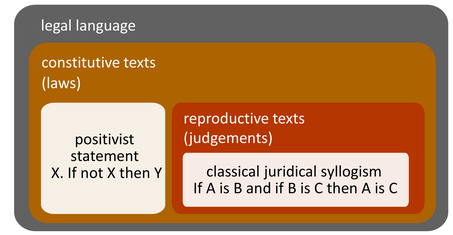
- MHG contains longer n-grams (maximal 12-grams versus maximal 8-grams in sCz)
- Bigrams are the most prominent units

Serving

- MHG seems to be more standardized (significantly more n-grams, longer n-grams) than sCz. However, there are some morphological and syntactic constraints like inflexion morphology and word order, which impacted the statistics in sCz.

Nota bene: This has also some theoretical implications on the n-gram notion. The question arises, whether one should include the semantic level in its description to cover the morphological and syntactic peculiarities of the Slavic languages; or consequently pursue the linear syntagmatic approach.

- It is not surprising that bigrams are usually the most frequent textual chunks, but their significance for the legal language is obviously connected with the text pattern and text type (cf. Kjær 2000):



- There are noticeable differences between sCz and Cz in the terminology and general pattern usage; which allows the hypothesis about the different ways of adoption of *ivs maideburgense* in the Bohemian and Slovak area.

Sächsische Akademie der Wissenschaften zu Leipzig

Sources
Kuchar, Rudolf 2009. Žilinská právna kniha. Preklad Magdeburského práva. Zápisy právnych úkonov Žilinských meštanov. Bratislava: VEDA – vydavatel'stvo Slovenskej akademie vied.
Piirainen, Ilpo Tapani (Hrsg.) 1972. Das Stadtrechtsbuch von Sillein. Einleitung, Edition und Glossar. Berlin, New York: de Gruyter. (=Quellen und Forschungen zur Sprach- und Kulturgeschichte der germanischen Völker. N. F. 46/170).
Práwa saszká: http://psp.cz/kps/knih/prawa/prawa.htm (accessed on 30.03.2015).

Selected references
Berger, Tilman 1997. Tschechen und Slowaken: Zum Scheitern einer gemeinsamen, tschechoslowakischen Schriftsprache. In: Hentschel, Gerd (Hrsg.). Über Muttersprachen und Vaterländer. Zur Entwicklung von Standardsprachen und Nationen in Europa. Frankfurt a. Main [usw.]: Peter Lang, S. 151-181.
Kjær, Anne Lise 1991. Phraseologische Wortverbindungen in der Rechtssprache? In: Palm, Christine (Hrsg.): Akten der Internationalen Tagung zur Germanistischen Phraseologieforschung: Aske, Schweden, 12. - 15. Juni 1990 / „EUROPHRAS 90". Stockholm: Almqvist och Wiksell Internat., S. 115-122.
Kjær, Anne Lise 1995. Vergleich von Unvergleichbarem. Zur kontrastiven Analyse unbestimmter Rechtsbegriffe. In: Kromann, Hans-Peder; Kjær, Anne Lise (Hrsg.): Von der Allgegenwart der Lexikologie: Kontrastive Lexikologie als Vorstufe zur zweisprachigen Lexikographie; Akten des Internationalen Werkstattgesprächs zur Kontrastiven Lexikologie, 29.-30.10.1994 in Kopenhagen. (= Lexicographica / Series maior ; 66). Tübingen: Niemeyer, S. 39-56.
Kjær, Anne Lise 2000. On the structure of legal knowledge. The importance of knowing legal rules for understanding legal texts. In: Lundquist, Lita (ed.): Language, text, and knowledge: Mental models of expert communication. (= Text, translation, computational processing ; 2). Berlin: Mouton de Gruyter, p. 127-161.
Kopaczyk, Joanna 2012. Long lexical bundles and standardisation in historical legal texts. In: Studia Anglica Posnaniensia 47, 2-3 (2012), S. 3-25.
Lieberwirth, Rolf 1986. Das sächsisch-magdeburgische Recht als Quelle osteuropäischer Rechtsordnungen. Berlin: Akademie-Verlag. (=Sitzungsberichte der Sächsischen Akademie der Wissenschaften zu Leipzig. Phil.-hist. Kl. Bd. 127. H. 1).
Reutter, Gerhard 1988. The language shift from Latin to Polish. Die Ablösung des Latein durch das Polnische in schriftlich geprägten Funktionsbereichen. In: Hill, Peter; Lehmann, Volkmar (eds): Standard language in the Slavic world. Papers on sociolinguistics by Hamburg Slavists. München: Sagner, p. 110-128.

Software
N-Gram Phrase Extractor: http://www.lextutor.ca/n_gram/ (accessed on 27.03.2015).
ParaConc, Collocate: http://www.paraconc.com/index.html (accessed on 20.10.2014).