

# Parallel historical corpora – a new method in standardization research?

Saxon Academy of Sciences Dr Marija Lazar

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



Stirbet aber iener auf den

dy clage get sein erbe

'If the one dies, who is

charged, his heir/heirs'

iener auf den dy clage get

'the one, who is charged'

(lit.: 'the one, upon whom

the claim comes')

Synthetic verbal forms

with inversion – bude li;

317 (12,11%)

81 (3,44%)

13 (0,59%)

2 (0,10%)

neda li; nema li

'if the one is /

does not give /

does not have'

n-grams Middle High German 'slovakized' Czech

Graph 1. The n-grams in the test corpus based on

• MHG contains longer n-grams (maximal 12-grams

Žilina Law Book (absolute and weighted)

494 (15,89%)

228 (7,81%)

100 (3,52%)

68 (2,4%)

• Bigrams are the most prominent units

Apples and oranges: sorting out

Long legal n-grams

ten [INT], na koho zaloba

gde, geho erbownik/erbowe

'the one, who is charged,

ten/onen/czlowiek, na

the one, who is charged'

(lit.: 'the one, upon

whom the claim comes')

A gestli by / A pakli by 'if'

Short n-grams shared with other domains

his heir/heirs'

Legal phrase

koho zaloba gde

Inversion

Boiling down

2-grams

3-grams

4-grams

5-grams

MHG contains more n-grams

versus maximal 8-grams in sCz)

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

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 <b>,</b> .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.



Kuchar, Rudolf 2009. Žilinská právna kniha. Preklad Magdeburského práva. Zápisy právnych úkonov Žilinských mešťanov. Bratislava: VEDA – vydavateľ 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).

n-grams: an indicator of standardization

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/)

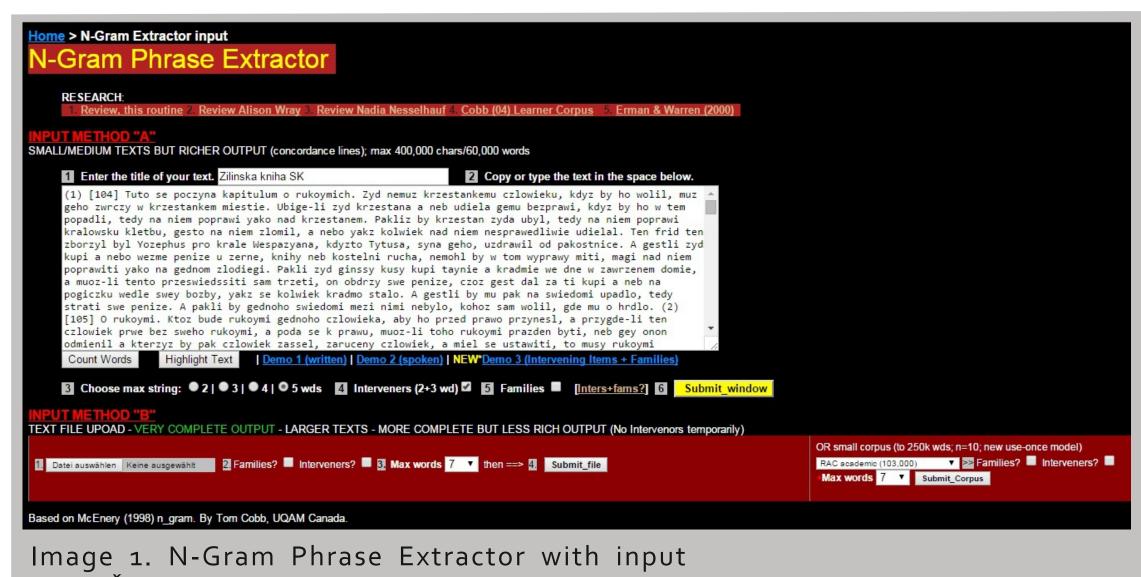
ParaConc ParaConc

www.korpus.sk NÁRODNÝ

www.korpus.cz ČESKÝ NÁRODNÍ KORPUS Vocabularies and lexical databases

Implementation Output output classification • exclusively legal n-grams as of n-gram length • n-grams shared with other domains classified n-grams • legal phraseological expressions • short n-grams (based on Kjær 1991) phraseological expressions shared (2-5-grams with the output comparison with other domains upload of the corpus interveners) between long n-grams data in the extractor, ((6+n)-grams with each language source text and translation completed terminology database with sCz interveners); (MHG versus sCz, Cz) separately check with ParaConc Slavic translations distributive analysis with explanation of the overlaps and the diachronic national ambiguity between sCz and Cz corpora for identification of specific legal n-grams

# Case study Harvesting



Input: Žilina Law Book test corpus (approx. 2700 tokens) Search parameters: 2-5-grams with interveners

Zilinska kniha SK (2720 wds) FAMILIES ARE 5-wd strings: 2072 Repeated phrases: 2 (0.10 %) Words involved: 21 / 2720 words (0.77%) 2072 total 5-wd strings: (repeats are <u>underlined</u>) 002 nez rychtarz ma hledieti na geho zbozi. 10 NUMBER S KTERYMI LIDMI czlowiek ma dlhy oswiedssiti. 003 ni zadneho czoz by k niemu przysluchal. 11 NUMBER TY DLHY KTEREZ z zalobu budu dobyty, ktery bu 004 toho chtiel przyti ma se za to boziti. 12 NUMBER O TEM PLATU kteryz bude nuznie prawem dobyty | 006 y muz tak drzeti yako prwe za swe dlhy. 16 NUMBER O ZALOBIE PRO dlhy. gyz slysste gestli obzalu 907 wiek czo by od toho chtiel plati dawat. **17 NUMBER O PRWE PLACENEM** dlhu. gyz slysste ma-ii geden 008 w geho zbozi ocz onen obzalowany gest. 25 NUMBER KTOZ SE PRZED prawem bozy. gyz słysste bozy-l 009 i a nepropadne osudu proti rychtarzowi. 26 NUMBER O TEM KTOZ se geden druhemu ma boziti. gyz s 010 tomu bozbu dati tak yako gde po prawie. 27 NUMBER KTOZ SE BOZY wedle swedomi. gyz slysste ma-li 011 ase ustawi na tu hodynu kdyz to slibil. 31 NUMBER O RUKOYMI O zwlassczni zalobu. kdyz geden czl 012 lo ta wina gest na tem rukoymi wyhrana. 33 NUMBER O DLUZE PO smrti toho czlowieka, take ma dano 013 dnu winu zadneho syrota bozby uczyniti. 35 NUMBER O BOZBIE SLIBENEY przed prawem. gyz słysste k 015 el pod swu ruku ocz onen byl obzalowan. 4 NUMBER O TEM CZLOWIEKU kterehoz za dlhy w ruce dadie. 016 d nedokona dodad bude wzdy geho zaklad. 5 NUMBER O DLZYCH KTEREZ platiti magi. tomu czlowieku k 3-wd strings: 2356 Repeated phrases: 81 (3.44 %) 017 ez w ten czas zgewnie gdu w tom prawie. 6 NUMBER O YATEHO CZLOWIEKA slibu. kazdeho czlowieka ya 018 slib w wiezeni ani bozba gest platena. 7 NUMBER O TEY BOZBIE kteraz nesskodi. kdyz se geden bo 019 geho newiernie k tomu slibu przysylil. 8 NUMBER O TEM KTOZ przed prawem platy pyta. ten czoz p 020 ebo za erb czoz by byl przyval pratiti. 9 NUMBER O TEM CZLOWIEKU kteremuz netrzeba pre dlhy ru 021 wedle wiedomi dobrych potcztiwych lidi A BRANI SE WEDLE SWIEDOMI on obdrzy swe penize welmi le 022 al. a czoz gest pak przed prawem slibil A BUDE PRZESWIEDSSEN TU WIECZ obdrzi se dwiema czlowie 023 w tom zachowal ten propadne rychtarowi. A BUDE-LI GEDNOMU CZLOWIEKU GEHO swiedomi rozluczene kt 024 u dobyty, ktery bude obzalowan pro dlhy A BUDE-LI KTOZ BY GEMU swiedssyl ma platiti we cztrnast 025 gedno rukoymi bude za gednoho czlowieka A BY GEI MIEL PRZED prawo ustawiti a przyssel by ten cz 026 yz słysste kdyz se gedna zena ma boziti A BYLA BY TAK NEMOCZNA ze by nemohla przed prawem stati 027 ginssi swu czeled s gidlem a s robotu. A CHCZE-LI GEHO PAK NA rzetiez wzyti nebo sputnat to mu 028 ti swiedomim lepe nez onem czo se prosy A CZOZ BY PAK CZLOWIEK w gistem swem kupecztwi nemiel r 029 przyti ktoz czo burguge to ma zaplatiti A CZOZ CZYNI TO MA ustawicznie drzeti, pakli by chtiel 030 czoz gest przed prawem neslibowal. A CZOZ GEST PAK PRZED prawem slibil a bude przeswiedsse 031 z onen swe kupeczstwi nan prowesti muze A DA GEMU ODPOWIED S hladku zalobu a pakli by se nebran 032 a przyssel by ten czlowiek przed prawo A DA SE SAM W prawo tym wedle swiedomi muoz oswoboditi 033 o obranugice miel sem gistotu na to rok A DEN A ZADNE PRAWO mi toho neodmluwalo a prawe dlzne z 034 wiek w gistem swem kupecztwi nemiel rok A DEN O TO MUSY przed prawem odmluwati w tu hodinu kdyz 035 to muoze uswiedssiti tedy gest swobodny A GEHO ERBOWE MAGI TOHO ustawiti a neb za toho umrleho 036 a ginssy slib netrzeba gemu toho plniti A GESSCZE PROTO OBDRZANY NA swey czti a muoze naò to do 037 w gedney cestie a neb na gednem mostie A GESTLI BY HONILI GYZNEHO a neb piessyho tedy wuoz ma
038 wey bozby yakz se kolwiek kradmo stalo. A GESTLI BY MU PAK na swiedomi upadlo tedy strati swe p

Image 2. The output from N-Gram Phrase Extractor

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). Гübingen: 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

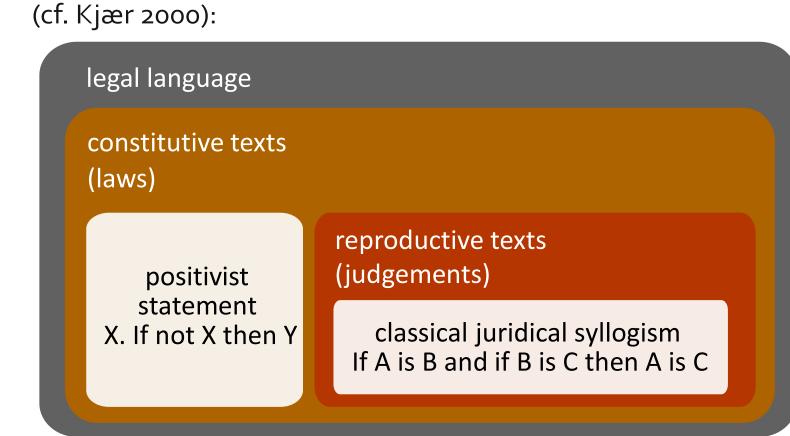
in the Slavic world. Papers on sociolinguistics by Hamburg Slavists. München: Sagner, p. 110-128.

# 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



• 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.

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 20.10.201

