First Results in Developing a Medieval Latin Language Charter Dictation System for the East–Central Europe Region

Péter Mihajlik[1,2], *Lili Szabó*[2], *Balázs Tarján*[1,2], *András Balog*[2] *and Krisztina Rábai*[3]

[1]Budapest University of Technology and Economics, Hungary
[2]THINKTech Research Center, Hungary
[3]University of Hradec Králové, Czech Republic
{mihajlik, lili, tarjan, balog}@thinktech.hu, rabaikri1@uhk.cz

Abstract

Latin had served as an official language across Europe from the Roman Empire until the 19th century. As a result, vast amount of Latin language historical documents (charters, account books) survived from the Middle Ages, waiting for recovery. In the digitization process, tremendous human efforts are needed for the transliteration of textual content, as the applicability of optical character recognition techniques is often limited. In the era of Digital Humanities our aim is to accelerate the transcription by using automatic speech recognition technology. We introduce the challenges and our initial results in developing a real-time, medieval Latin language LVCSR dictation system for East–Central Europe (ECE). In this region, the pronunciation and usage of medieval Latin is considered to be roughly uniform. At this phase of the research, therefore, Latin speech data was not collected for acoustic model training but only for test purposes – from a selection of ECE countries. Our experimental results, however, suggest that ECE Latin varies significantly depending on the primary national language on both acoustic–phonetic and grammatical levels. On the other hand, unexpectedly low word error rates are obtained for several speakers whose native language is completely uncovered by the applied training data.

Index Terms: speech recognition, real-time LVCSR, medieval Latin, East–Central Europe, cross-lingual/multi-lingual acoustic and lexical modeling

1. Introduction

In this paper, our aim is to investigate the feasibility of developing a Latin language Large Vocabulary Continuous Speech Recognition (LVCSR) system for dictation of historical documents. Such a technical aid would be highly beneficial for historians and philologists in general – and specifically in the digitization process of medieval charters. In the Middle Ages, Latin as a spoken language changed substantially from its classical state – dialects developed into independent languages, such as Spanish, French, Italian and Romanian. At the same time, as an official language in Europe, it had changed more softly, and this dialect will be referred to as 'medieval Latin' in the rest of the paper. Obviously, national languages had a certain influence on medieval Latin but the pronunciation and grammar is considered more or less consistent in such a geographical region as East–Central Europe (ECE) [1–2]. Therefore we decided to start experimentation with charter texts written originally in the ECE region and read aloud by native speakers of the same geographical area. Regarding similar researches, only loosely related works could be found [3-6]. To the best of our knowledge, no prior publication, research or commercial application is dealing with Latin LVCSR – not to mention the dictation in the medieval dialect of Latin in the ECE region.

2. Data acquisition

State-of-the-art speech dictation systems are trained on a large amount of task specific text and speech data. However, as Latin is no longer used for everyday communication and the texts has to be extracted from hard-to-access, medieval documents; we face real difficulties in getting relevant data.

2.1. Text

2.1.1. Collection issues, corpora

There are plenty of Latin language documents available, but only a small fraction is digitized. Many of the digitized historical documents are available through the web, but typically only in image formats, without the text content extracted. Even if the text is available electronically, Latin is frequently mixed up with national languages and the metadata is not always adequate for automatic extraction of the Latin text. In our case the following conditions narrowed the search space even further: the data had to be originated from the ECE region and from the Middle Ages. Moreover, as our target application is charter dictation, a relevant amount of in domain – charter or diploma – text was required. Finally, the following language resources were collected.

Training set: two different sources were used to build the language model training corpus. The first one, referred to as '*Monasterium*', is a charter archive [7] ranging from AD 1000 to 1524 consisting of charters comprising 500k tokens. Since these charters were created in the late Hungarian Kingdom, they are presumably influenced by the local Hungarian language. The other corpus named '*LatinLibrary*' consists of medieval Latin literary and historical texts [8] comprising 1.3 million tokens.

Test set: using independent sources, 3 charters originating from Czech, Hungarian and Polish regions were selected for the evaluation of the language model (LM) and for readings for speech recognition tests. There was no overlap between any training and test sets on the sentence level.

2.1.2. Preprocessing

Even contemporary texts may contain several typos and the spelling of foreign names is often not uniform resulting in a degradation in language modeling. In the Middle Ages the situation was even more peculiar as spelling was not standardized.

An interesting feature of the corpora is that they contain a significant number of spelling variants, e.g. "iuris" and "juris". To detect the spelling variants we took all items in the pronunciation dictionary (see Section 3.1.2.) whose pronunciation were identical, and used context and expert knowledge to decide whether the equivalent pronunciations mean spelling variants or homophones. Where the decision was that they are spelling variants, the less frequent one was replaced by the more frequent one.

Resolving spelling variants resulted in a more consistent corpus in terms of overall perplexity (PP) reduction from 775 to 672 and OOV rate decrease from 4.3% to 3.5% (see next section for LM details).

2.1.3. Evaluation

To assess the relevance of LM training texts, perplexity and OOV rate measurements were performed on the test charter texts as Table 1 shows. Modified interpolated Kneser–Ney smoothing [9] was applied on the word 3–gram language models built from the Monasterium and LatinLibrary corpora after preprocessing. The LM's of the two data set were linearly interpolated and used later in all speech recognition experiments.

Table 1: *Perplexity and OOV rate evaluation of LM's over the origination regions of Latin test texts.*

LM Corpus	**Latin Test Text** PP/OOV rate			
Vocab. size	CZ	PL	HU	∑
Monasterium	3130	551	82	479
35k	18.3%	11.8%	0.9%	10.5%
LatinLibrary	2305	3266	3549	2996
115k	5.5%	7.8%	1.6%	9.7%
Interpolated	2288	924	82	672
133k	5.5%	3.9%	0.9%	3.5%

As can be seen in Table 1, the Monasterium corpus and Hungarian test charter are well matched. The "general" medieval LatinLibrary corpus is indeed very different from the test charters but its addition to the final, interpolated LM decreased the speech recognition error rates constantly (results not discussed here).

2.2. Speech

Although our experiences show a certain difference between dictational and read speech, there was no other option than to start with read speech, for *test* purposes. Medieval Latin speakers from several ECE countries were invited to read aloud the 3 previously described charter texts. Recording conditions were accurately controlled: close–talking microphones, quiet, non reverberant acoustic environment, fluent, flawless speech and at least 16kHz, 16 bit (linear PCM) encoding. No instructions were given to the speakers on how to pronounce medieval Latin words or named entities. At the end, a data set was collected with following native language speakers: *Czech, Slovakian, Polish, Lithuanian* and *Hungarian.* The overall length of the recorded test speech was around 35 minutes.

Oral and lingual test data can be accessed and contributed on the project website.[1]

[1] http://medilatin.speechtex.com/

3. Speech Recognition Approaches

The fundamental problem was that no Latin speech data was directly applicable for acoustic model training. Therefore, we decided to apply our readily available ECE speech corpora for that purpose – as a reasonable choice considering the nationalities of the test speakers. In the following, we introduce the various acoustic modeling approaches applied.

3.1. Hungarian based acoustic modeling

3.1.1. Phoneme acoustic model

Phoneme acoustic model was trained in a classical way: numerous manually transcribed Hungarian speech corpora had been added to the training set. The two main sources of data were the Speecon database [10] and a broadcast speech database we had collected for a previous research [11]. Altogether, a Hungarian language database with a length of more than 500 hours from several thousand speakers was obtained. G2P (Grapheme to Phoneme) mapping between orthographic transcriptions and Hungarian phonemes was performed utilizing language specific rules and pronunciation dictionaries. The phonetic transcriptions were used for training a context dependent DNN (Deep Neural Network) acoustic model. For further technical details, see Section 4.

3.1.2. Pronunciation model for Latin

In the ECE pronunciation system graphemes mostly correspond to individual phonemes with some exceptions. We applied context independent digraph transcriptions such as 'qu' → /kv/, 'ph' → /f/, 'ae' → /e/ and 'oe' → /ø/. As for the context sensitive rules ch became /h/ in the presence of a preceding vowel and /k/ otherwise. The consonant 'c' became /ts/ if followed by a palatal (front) vowel, and /k/ otherwise. The letter sequence 'ti' became /tsi/ if followed by a vowel and not preceded by 's', 't' or 'x', and /ti/ otherwise. Similarly 'gu' became /gv/ when followed by a vowel. Other than the above main rules, all geminate consonants became de–geminated (e.g. litteris → /literis/). Finally, the Hungarian phoneme inventory set applied for Latin consisted of 24 elements.

3.2. Czech based acoustic modeling

In addition to the Hungarian Speecon database the Czech Speecon [12] was also available for this research. The phoneme acoustic models were trained in the same way as in the case of Hungarian. We used the Czech phoneme inventory set definition and dictionary included to the database. G2P from Latin text to Czech phonemes were performed based on the same principles as in Section 3.1.2. The Czech phoneme inventory set used in Latin speech recognition experiments consisted of 26 items.

3.3. Polish and Romanian based acoustic modeling

A substantial amount of Polish and Romanian speech data from television broadcast news were collected and transcribed in our laboratories [13] that could be utilized in this project. The challenge was that no G2P knowledge was readily available for these languages and therefore grapheme based acoustic models could be used only. As in Romanian and in Polish the orthography is nearly phonemic, thus we decided to apply the grapheme acoustic models in the following two ways.

3.3.1. *Direct grapheme modeling*

Even though basic Latin graphemes denote often different phonemes compared to Romanian or Polish graphemes, a direct, G2P free approach was tested. The graphemes missing from the Latin grapheme set were simply ignored, otherwise the acoustic model of the corresponding native grapheme was used.

3.3.2. *Graphemes as phonemes*

In this attempt, Latin graphemes were indeed mapped to Hungarian phonemes. Hungarian phonemes then were mapped again to the corresponding Romanian or Polish grapheme acoustic models in a 'best effort' manner. So, in this approach, Hungarian phonemes act as an intermediary layer between Latin words and grapheme acoustic models of Polish or Romanian.

3.4. Unified simplified grapheme (USG) acoustic modeling

Probably, it would have been a good idea to develop a merged phoneme inventory set [6] for all the four ECE language speech data – but given the lack of G2P conversion rules for Romanian and Polish, another solution was required. Using graphemes instead of phonemes as acoustic units is also an issue since these 4 languages – though each one's alphabet is based on Latin – use significantly different character sets.

The technique we propose is referred to as Unified Simplified Grapheme (USG) approach described in the following. Without any G2P mapping, the transcription texts of the four–language speech training data were sliced into letters and all the variants of the original Latin characters are mapped back to their base form, e.g.: ţ → t, á → a, ž → z. As a result, a unified simplified grapheme inventory set with a size of 26 (reduced from 58) was produced, compatible with medieval Latin. The USG units were then used as acoustic model units in the training. The background assumption behind this seemingly oversimplified approach is that the modified Latin characters of the ECE languages should have some relation to their – may be medieval Latin – ancestor.

To avoid the over weighting of Hungarian language data in the USG acoustic model only the broadcast news parts of Hungarian database collection were kept in the training.

4. Experimental Results

4.1. Experimental system

In this section we introduce the techniques used to train our real–time, Latin dictation system. Further details about the applied LVCSR technology can be found in [14–15].

4.1.1. *Acoustic model training*

Phoneme (P), Grapheme (G) and USG acoustic models were trained on the same way. Mel-Frequency Cepstrum + Energy features were used along with Linear Discriminant Analysis (LDA) + Maximum Likelihood Linear Transformation (MLLT), with a splice context of ±4 frames, 10ms of frame shift. 9x40 dimension, spliced up feature vectors served as input to the feed–forward, 6 hidden–layer neural network with p–norm [16] activation function. Prior to DNN training, a Gauss Mixture Model (GMM) pre-training was performed and Clustering and Regression Tree (CART) [16] was applied to obtain across–word context dependent shared state phone (or 'graph') models and their time alignment. The number of senones (and so the size of the DNN softmax output layer) was between 7,000 and 11,000 depending on the nature of the training data. The size of the hidden layers was kept constantly on 2,000. A minibatch size of 512, a initial learning rate of 0.1 and final learning rate of 0.01 was applied in 20 epochs using the KALDI toolkit [16].

4.1.2. *WFST recognition network construction & decoding*

The interpolated LM described in Section 2.1 was composed with the appropriate phoneme or grapheme based pronunciation dictionaries in the Weighted finite State Transducer (WFST) framework [17]. After the usual optimization processes the WFST recognition network translates between generalized triphones (or trigraphs) and words. The mapping of triphones (trigraphs) to senones is performed in the VoXerver WFST–HMM–DNN decoder. Approximately the same speed, faster than real–time decoding was performed in all experiments.

4.2. Results & discussion

4.2.1. *Overall results*

The overall speech recognition result of the various acoustic and pronunciation modeling approaches are introduced in Table 2.

As expected, one of the best result is obtained when using the Latin grapheme to Hungarian phoneme mapping method coupled with the largest acoustic training database. The application of task specific Latin pronunciation rules improved the performance of the Polish and Romanian grapheme–based systems as well. However, the word error rates of all monolingual grapheme acoustic models are elevated, roughly two times higher than the best WER. In addition to that, the Czech phoneme based result is below expectations – possible explanations might be the relatively small size and the different nature of the acoustic and textual data. Grapheme based Hungarian and Czech approaches were also applied but gave significantly worse results, not discussed here.

On the other hand, the USG acoustic modeling approach performs beyond expectations. Without any explicit knowledge of the test language, the result is competitive with the Hungarian phoneme based setting. Note, that even if the USG system was trained on four East–Central European languages speech data, the overall size of the applied database is less than the half of the Hungarian one.

Table 2: *Overall WER results on medieval Latin charter dictation*

Acoustic model			**WFST G2P [Latin to ...]**	**WER [%]**
Language	**Size [hour]**	**Unit**		
HU	567	P	HU	26.9
CZ	76	P	CZ	46.4
PL	31	G	–	59.8
PL	31	G	HU	50.8
RO	35	G	–	60.5
RO	35	G	HU	48.0
HU+ CZ+ PL+ RO	112 +76 +31 +35	USG	–	26.2

Table 3: *Detailed WER [%] of **Hungarian phoneme based** medieval Latin dictation – over the origination regions of Latin test texts and the native language of speakers.*

	Latin Test Text			
Speaker	CZ	PL	HU	∑
CZ	19.1	27.5	6.4	17.8
SK	20.4	22.9	9.1	17.8
PL	29.0	40.4	15.5	28.3
LT	49.3	51.4	20.9	41.5
HU	25.0	20.2	25.5	23.7
∑	30.2	34.5	14.9	*26.9*

Table 4: *Detailed WER [%] of **4 ECE languages USG based** medieval Latin dictation – over the origination regions of Latin test texts and the native language of speakers.*

	Latin Test Text			
Speaker	CZ	PL	HU	∑
CZ	20.1	33.0	11.8	21.4
SK	14.5	24.8	12.7	17.0
PL	23.0	33.9	10.0	22.4
LT	48.0	53.2	17.3	40.4
HU	20.4	25.7	14.6	20.2
∑	27.7	36.7	13.6	*26.2*

4.2.2. Detailed WER analysis

Under such heterogeneous test conditions – where the 3 parts of the test data consist of medieval Latin texts influenced by 3 different local languages with test speakers whose native languages cover 5 different ECE languages – a detailed diagnostic analysis of WER is necessary. Therefore, recognition results – organized in a 3x5 matrix of test text and speaker nationalities – have been investigated. The detailed results of the Hungarian phoneme and the USG approaches are of particular interest, discussed on Table 3 and 4.

First of all, it can be clearly seen that WER results on the Hungarian origin test text, in average, are less than half of the ones measured on Polish and Czech origin charter texts. This is fully explained by the fact that the LM training text is matched better (in terms of perplexity and OOV rate) to Hungarian text than to the other ones.

What is more, the results of Czech and Slovak speakers on this Hungarian text are outstanding even on an absolute scale – at least in case of the Hungarian phoneme based system. Unexpectedly, these results are significantly better than the corresponding one of the Hungarian speakers. By inspecting the test recordings, we found the following phenomena that can explain these strange results. First, the Latin to Hungarian G2P was developed based on written national pronunciation rules similar throughout the ECE region – but we could not find an exact and full description of Latin pronunciation rule set. Therefore the implementation may contain some simplifications or minor modifications. So, possibly the actual Czech an Slovak pronunciation is closer to the implemented G2P mapping than the actual Hungarian one. An auxiliary explanation is that – because of the geographical proximity and the shared history – acoustic-phonetics of Czech and Slovak are somewhat similar to the Hungarian (even if the grammar and morphology is rather different), as opposed to Lithuanian or Polish. Finally, named entities are often read according to the native language pronunciation rules which differs form the Latin ones resulting in a recognition error increase in the case of a native speaker – similar phenomenon can be observed in the Polish case.

By comparing the results in Table 3 and 4, it can be noticed that the USG approach gives not only the best overall result but also its partial results are more balanced – mostly Lithuanian is the sole outlier. This can be explained by the more data driven way of G2P modeling in the case of the USG system that is less sensitive to pronunciation variations than a rule based approach. Again, the slight degradation in the Czech speakers' results and the small improvement in the Hungarian speakers' one are against expectations. This phenomenon might be explained by the better modeling ability of the USG approach regarding long consonants (geminates) which are expressed more in Hungarian pronunciation than in Slavic languages. Verified by Wilcoxon signed-rank test, the overall difference between the two best approaches is not statistically significant.

Based on the detailed recognition results, it can be concluded that acoustic and pronunciation levels of Medieval Latin depend significantly on the speaker's native language even in the ECE region.

5. Conclusions

We introduced our first results of a medieval Latin speech recognition system tailored for East–Central Europe charter dictation. No Latin speech data was used for training the acoustic models and gathering relevant text data was also challenging. It was shown that despite the difficulties, the development of a real-time medieval Latin LVCSR system is not necessarily an impossible mission – at least for Czech and Slovak speakers and in case of matched test text where a WER of less than 10% was measured. Apart from a conventional, Hungarian phoneme based speech recognition approach, a unified simplified grapheme based technique was also introduced and applied in the training on a four-language speech data. The latter technique provided competitive results practically without any explicit knowledge on the grapheme to phoneme rules of the training and test languages.

As for future work, a plenty of research directions are open. Obviously, the collection a substantial amount of medieval Latin speech for training purposes would be highly beneficial, though not trivial. An other direction could be the development of more elaborated Latin G2P rules customized for speaker or native language – used together with a global phoneme inventory set. The application of advanced adaptation techniques should also be investigated. Finally, based on the unified simplified grapheme approach, the extension of speech corpora composed of ECE national languages – in terms of further languages and size – can yield in promising results.

6. Acknowledgements

The authors would like to thank the Speakers for kindly providing the test recordings as well as SpeechTex Inc. and Tibor Fegyó for their support and for providing the speech recognition engine designed for dictation.

7. References

[1] K. C. Sidwel, "Reading Medieval Latin" *Cambridge Unversity Press*, 1995.

[2] P. Ford, J. Bloemendal, C. Fantazzi: "Brill's Encyclopeadia of the Neo-Latin World: Macropedia" *Leiden-Boston*, 2014.

[3] B. McGillivray, A. Kilgarriff, "Tools for historical corpus research, and a corpus of Latin", in *New methods in historical corpora*, pp. 247–256, 2013.

[4] A. Clareborn, P. Cullhed, J. Edlund, "Transcription by dictation" in *Digital Humanities in the Nordic countries*, 2016.

[5] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under–resourced languages: A survey," *Speech Communication,* vol. 56, pp. 85 – 100, 2014.

[6] T. Schultz, A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition", in *Speech Communication 35 (1)*, pp. 31–51, 2001.

[7] Monasterium: http://monasterium.net/mom/HU-PBFL/PannHOSB/fond?block=15#ch1

[8] LatinLibrary: http://www.thelatinlibrary.com/medieval.html

[9] A. Stolcke, "SRILM – an extensible language modeling toolkit", in *Proc. Intl. Conf. on Spoken Language Processing*, pp. 901–904, Denver, 2002.

[10] Hungarian Speecon database, http://catalog.elra.info/product_info.php?products_id=1094, 2003.

[11] A. Varga, B. Tarjan, Z. Tobler, G. Szaszak, T. Fegyo, C. Bordas , P. Mihajlik, "Automatic Close Captioning for Live Hungarian Television Broadcast Speech: A Fast and Resource-Efficient Approach", in *LNAI 9319: pp. 105–112. 17th International Conference on Speech and Computer (SPECOM)*. Athens, Greece, 2015.

[12] Czech Speecon database, http://catalog.elra.info/product_info.php?products_id=1095, 2004.

[13] B. Tarjan, T. Mozsolics, A. Balog, D. Halmos, T. Fegyo, P. Mihajlik, "Broadcast News Transcription in Central-East European Languages", *3rd IEEE International Conference on Cognitive Infocommunications. Košice, Slovakia*, pp. 59–64, 2012.

[14] L. Szabo, B. Tarjan, P. Mihajlik, "Writing with Speech: A Qualitative User Evaluation Study", in: *5th IEEE International Conference on Cognitive Infocommunications: CogInfoCom 2014: Proceedings, Vietri sul Mare, Italy*, pp. 355–360, 2014.

[15] P. Mihajlik, Z. Tuske, B. Tarjan, B. Nemeth, T. Fegyo, "Improved Recognition of Spontaneous Hungarian Speech: Morphological and Acoustic Modeling Techniques for a Less Resourced Task", *IEEE Transactions on Audio Speech and Language Processing* 18:(6) pp. 1588–1600, 2010.

[16] D. Povey, A. Ghoshal et al., "The Kaldi Speech Recognition Toolkit," *in Proceedings ASRU,* 2011.

[17] M. Mohri, F. Pereira and M. Riley, "Weighted Finite-State Transducers in Speech Recognition", *Computer Speech and Language, 16(1)*, pp. 69–88, 2002.