EPiC Series in Language and Linguistics
Volume 4, 2019, Pages 113–124
Proceedings of Third Workshop "Computational linguistics and language science"
EPiC Language and Linguistics

Lemmatisation for under-resourced languages with sequence-to-sequence learning: A case of Early Irish

Oksana Dereza[1,2]

[1] National Research University “Higher School of Economics”, Moscow, Russia
https://www.hse.ru/en/staff/odereza
odereza@hse.ru
[2] Lomonosov Moscow State University, Moscow, Russia

Abstract

Lemmatisation, which is one of the most important stages of text preprocessing, consists in grouping the inflected forms of a word together so they can be analysed as a single item, identified by the word’s lemma, or dictionary form. It is not a very complicated task for languages such as English, where a paradigm consists of a few forms close in spelling; but when it comes to morphologically rich languages, such as Russian, Hungarian or Irish, lemmatisation becomes more challenging. However, this task is often considered solved for most resource-rich modern languages irregardless of their morphological type. The situation is dramatically different for ancient languages characterised not only by a rich inflectional system, but also by a high level of orthographic variation, and, what is more important, a very little amount of available data. These factors make automatic morphological analysis of historical language data an underrepresented field in comparison to other NLP tasks. This work describes a case of creating an Early Irish lemmatiser with a character-level sequence-to-sequence learning method that proves efficient to overcome data scarcity. A simple character-level sequence-to-sequence model trained during 34,000 iterations reached the accuracy score of 99.2 % for known words and 64.9 % for unknown words on a rather small corpus of 83,155 samples. It outperforms both the baseline and the rule-based model described in [21] and [76] and meets the results of other systems working with historical data.

1 Introduction

One of the biggest problems one faces working on NLP tools for under-resourced languages is the lack of data. It is widely known that in machine learning the quality of a model largely depends on the size of the training corpus. The situation is even more dramatic when it comes to ancient and medieval texts, since historical language data is not only sparse, but also very inconsistent.

Lemmatisation, which is one of the most important stages of text preprocessing, consists in grouping the inflected forms of a word together so they can be analysed as a single item, identified by the word’s lemma, or dictionary form. It is not a very complicated task for languages such as English, where a paradigm consists of a few forms close in spelling; but when

G. Wohlgenannt, R. von Waldenfels, S. Toldova, E. Rakhilina, D. Paperno, O. Lyashevskaya, N. Loukachevitch, S.O. Kuznetsov, O. Kultepina, D. Ilvovsky, B. Galitsky, E. Artemova and E. Bolshakova (eds.), CLLS 2018 (EPiC Series in Language and Linguistics, vol. 4), pp. 113–124

it comes to morphologically rich languages, such as Russian, Hungarian or Irish, lemmatisation becomes more challenging. However, this task is often considered solved for most resource-rich modern languages irregardless of their morphological type. The situation is dramatically different for ancient languages characterised not only by a rich inflectional system, but also by a high level of orthographic variation.

Old and Middle Irish, often described together as “Early Irish", is a language with an extremely complicated inflectional system and a high level of orthographical variation. It means that an average number of forms for each lemma in Early Irish will be substantially bigger than in many other European languages. Therefore, a training corpus for a task of lemmatisation in this case must be substantially bigger as well for any machine learning algorithm to work. The problem is, there are no publicly available annotated corpora of Early Irish, except POMIC [41], which is represented as a bunch of parse trees in PSD format, thus being not a very suitable source of data for machine learning.

Is there any solution except manually annotating all the digitised texts first, and then building ML-based NLP tools, or opting for rule-based systems? It seems like going down from word-level to character-level and using sequence-to-sequence learning might help. If we reformulate the lemmatisation task as taking a sequence of characters (form) as input and generating another sequence of characters (lemma), we can forget about tens of verbal and nominal inflection classes, let alone spelling variation. Moreover, this approach allows us to use the Dictionary of the Irish Language [68] as source of data.

This work describes a case of creating an Early Irish lemmatiser with a character-level sequence-to-sequence learning method that proves efficient to overcome data scarcity.

2 Related Works

The problem of NLP for historical languages first arose in the last quarter of the XX$^{\text{th}}$ century in regard to Ancient Greek [48], Sanskrit [71, 31] and Latin [44, 49] and for a long time was confined to these languages. As more and more medieval manuscripts were being digitised, there appeared a number of works dedicated to spelling variation in historical corpora, its normalisation and further linguistic processing for Early Modern English [5, 6], Old French [66], Old Swedish [10], Early New High German [9], historical Portuguese [29, 56, 27], historical Slovene [58], Middle Welsh [46] and Middle Dutch [36, 37]. Historical data processing in general has been surveyed in a substantial monograph [53] and several articles [25, 52]. Apart from corpus studies, there have emerged several open-source tools for historical language processing, such as a Classical Language Toolkit[1] [34], which offers NLP support for the languages of Ancient, Classical, and Medieval Eurasia. For the moment, only Greek and Latin functionality in CLTK includes lemmatisation.

Lemmatisation has also been an active area of research in computational linguistics, especially for morphologically rich languages [19, 20, 43, 14, 15, 63, 28, 69]. There are two major approaches to lemmatisation, a rule-based approach and a statistical one. The rule-based approach, which requires much manual intervention but yield very good results due to being language-specific, is widely used, examples being Swedish [17], Icelandic [32], Czech [35], Slovene [54], German [51], Hindi [50], Arabic [3, 24] and many other languages. A classical work on automatic morphological analysis of Ancient Greek describes a stem lexicon, where each stem is marked with inflectional class, and a list of pseudo-suffixes needed to restore these stems to lemmas [48]. A Latin lemmatiser from the aforementioned Python library CLTK also

[1] http://docs.cltk.org/en/latest/

uses stem and suffix lexicons. The best morphological analyser for Russian, Mystem, is based on Zalizniak grammatical dictionary [77]. This dictionary contains a detailed description of ca. 100,000 words that includes their inflectional classes. Mystem analyses unknown words by comparing them to the closest words in its lexicon. The 'closeness' is computed using the built-in suffix list [61]. A morphological analyser of modern Irish used in New Corpus of Ireland is based on finite-state transducers and described in [22] and [38].

Statistical approach to lemmatisation is computationally expensive and requires a large annotated corpus to train a model, especially when one deals with a complex inflectional system. Nevertheless, there are a few statistical parsers that achieve excellent results. Morfette, which was developed specially for fusional and agglutinative languages, simultaneously learns lemmas and PoS-tags using maximum entropy classifiers. It does not need hard-coded lists of stems and suffixes and derives lemma classes itself from the working corpus [16]. It shows over 97 % lemmatisation accuracy for seen words and over 75 % accuracy for unseen words on Romanian, Spanish and Polish data. Another joint lemmatisation and PoS-tagging system, Lemming, achieves more than 93-98 % for both known and unknown words on Czech, German, Spanish and Hungaian datasets [47]. Now there are models available for more than 15 languages, including Basque, Hebrew, Korean, Estonian, French and Arabic[2]. Unfortunately, it is almost impossible to directly compare the performance of rule-based and statistical-based systems for the same language described in different works due to the discrepancy of training datasets and the absence of evaluation results for some of the models.

Recently, neural networks also started being used for lemmatisation. A system combining convolutional architecture that models orthography with distributional word embeddings that represent lexical context was successfully implemented by [37] to lemmatise Middle Dutch data. The authors obtained 94-97 % accuracy for known words and 45-59 % accuracy for unknown words on four different datasets.

3 Data

3.1 Sources

One of the most difficult problems one faces working on NLP tools for ancient languages is the lack of data. The quality of a machine learning model is widely known to depend upon the size of the training corpus. The only publicly available annotated corpus of Early Irish is POMIC [41], but it is not a very suitable source of data for machine learning because it is represented as parse trees in PSD format. Another substantial resource is the electronic edition of the Dictionary of the Irish Language[3] [68]. The DIL is a historical dictionary of Irish, which covers Old and Middle Irish periods. Each of 43,345 entries consists of a headword (lemma), a list of forms including different spellings and compounds and examples of use with a reference to source text.

However, the list of forms cited in the DIL is incomplete; apart from that, some of the forms are contracted: for example, the list of forms for *cruimther* 'priest' is represented in the dictionary as -ir, which stands for *cruimthir*, and the list of forms for *carpat* 'chariot' looks like *cairpthiu, -thib, -tiu, -tib*, which has to be read as *cairpthiu, caipthib, cairptiu, cairptib*. Words can be abbreviated in many different ways, which is a consequence of the fact that there were many scholars who contributed to the DIL throughout 1913-1976, and each of them used his

[2] http://cistern.cis.lmu.de/marmot/models/CURRENT/
[3] http://dil.ie

Table 1: Contracted, restored and missing forms and spellings from the DIL

DIL	**Restored**	**Missing**
carpat, cairpthiu, -thib, -tiu, -tib	carpat, cairpthiu, caipthib, cairptiu, cairptib	carbad, carbat, carbait, carpait, carput, carpti...
carat(r)as	caratas, caratras	caratrad, caradras, caradrus, caradruis, caratrais...
cruimther, -ir	cruimther, cruimthir	cruimter, crumther, cruimthear, crumper, crumpir, cromthar, crumthirech
anmothaig[thig]e	anmothaige, anmothige	anmothaigthech, anmotuighe...
aball, a.	aball	abhull, aboll, ubull, abaill, abla, abhla, ubla, ubhaill...

own notation, as preserved in the digital edition. Some common types of contractions are listed in Table 1.

Still, the DIL is the best source of data for training a lemmatiser. The electronic edition of the DIL [68] was used to compile a training corpus of 83,155 unique form-lemma pairs, extracted from HTML files and restored to their full forms when necessary. These samples were then shuffled and split into training, validation and test sets, the former two being 5,000 samples each. One has to bear in mind, that this amount of training data is still insufficient for getting extremely good results in lemmatisation for a language as morphologically complex and orthographically inconsistent as Early Irish.

3.2 Morphology and Orthography

Old Irish is a fusional language with an elaborate system of verbal and nominal inflexion, comparable to Ancient Greek and Sanskrit in its complexity. In Celtic languages, there are two ways to encode morphological information in a word form, which often occur together: regular endings and grammaticalised phonetic changes in the beginning of the word called 'initial mutations'. It means that the first sound of a word can change under specific grammatical conditions, for example, the word *céile* 'servant' with a definite article in nominative plural will take a form *ind chéili* 'the servants', where the first stop [k] mutated into fricative [x]. This type of mutation is called lenition, and in this particular case it shows the presence of a definite article in nominative plural masculine, while the ending *-i* means that the noun itself is in nominative plural. There are four types of initial mutations in Early Irish: lenition, eclipsis, t-prothesis and h-prothesis. I will not expand on how exactly they affect consonants and vowels and when they occur, because it is not relevant for the task. I have to mention though, that both in Old and Middle Irish mutations were inconsistently marked in writing, and the orthography on the whole involves much variation.Tables 2 and 3 show various spellings of mutated vowels and consonants I encountered in my data.

There are several other orthographic features that increase a number of possible forms for a single lemma:

Table 2: Mutated consonant spellings

Original	b	c	d	f	g	l	m	n	p	r	s	t
Mutated	bh	ch	dh	fh	g	ll	mh	nn	ph	rr	sh	th
	mb	gc	nd	ḟ	ng	l-l	mm		bp		ṡ	dt
		cc		ḟh			m-m				ss	
				bhf							ts	
											s-s	

Table 3: Mutated vowel spellings

Original	a	á	e	é	i	í	o	ó	u	ú
Mutated	ha	há	he	hé	hi	hí	ho	hó	hu	hú
	na	ná	ne	né	ni	ní	no	nó	nu	nú
	n-a	n-á	n-e	n-é	n-i	n-í	n-o	n-ó	n-u	n-ú
	t-a	t-á	t-e	t-é	t-i	t-í	t-o	t-ó	t-u	t-ú

- inconsistent use of length marks;
- in later texts there appear mute vowels that indicate the neighbouring consonant's quality;
- complex verb forms can be spelled either with or without a hyphen or a whitespace.

Moreover, in Old and Middle Irish objective pronouns and relative particles are incorporated into a verb between the preverb and the root: cf. *caraid* 'he / she / it loves' and *rob-car-si* 'she has loved you', where *ro-* is a perfective particle, *-b-* is an infixed pronoun for 2^{nd} person plural object, and *-si* is an emphatic suffixed pronoun 3^{d} person singular feminine. The presence of a preverb with dependent forms triggers a shift in stress, which causes complex morphophonological changes and often produces a number of very differently looking forms in a verbal paradigm, particularly in the case of compound verbs, cf. *do-beir* 'gives, brings' and *ní tab(a)ir* 'does not give, bring'. Table 4 illustrates the variety of Early Irish verbal forms through the example of *do-beir*.

I should also mention, that the DIL is not strictly grammatical in the following assumptions, and so are the models trained on it:

- verbal forms with infixed pronouns are lemmatised as verbal forms without a pronoun (*notbéra* 'will bring you' > *beirid* 'brings');
- compound forms of a preposition and a definite article are lemmatised as prepositions without an article (*isin* 'in + DET' > *i* 'in');
- prepositional pronouns are lemmatised as prepositions (*indtib* 'in them' > *i* 'in');
- emphatic suffixed pronouns (*-som, -siu, -si, -sa* etc.) are lemmatised as independent personal pronouns.

Table 4: Some forms of the verb ‘do-beir’

Form	Deutero-tonic	Prototonic (after preverb)	Translation
INDIC PRES 3SG	do-beir	(ní) thabair	‘does (not) give / bring’
SUBJ PRES 3SG	do-bera	(ní) thaibrea	‘if does (not) give / bring’
PRET 3SG	do-bert	(ní) thubart	‘did (not) give / bring’
FUT 3SG	do-béra	(ní) thibéra	‘will (not) give / bring’
PERF 3SG	do-rat	(ní) tharat	‘did (not) give’
PERF2 3SG	do-uic	(ní) thuicc	‘did (not) bring’

Table 5: Character-to-character model mistakes

Form	Real lemma	Predicted lemma
ar-com-icc	ar-cóemsat	ar-coimcin
dáirfiniu	dáirine	dáirfinu
folortadh	folortad	folortaid
fris-tasgat	fris-tasgat	fris-taig
ithear	ithir	íthra
n-etarcnaigedar	etargnaigidir	etarncaigedar
t-iarrath	íarrath	dírarth

4 Experiment and Evaluation

A character-to-character model was trained during 34,000 iterations, but reached minimum loss and maximum accuracy of 69.8 % on a validation set after 10,000 iterations. When the training set accuracy reached its maximum, the validation set accuracy dropped to 64.9 %; on the test set the model achieved 63.9 %, , as shown in Figure 1. These results are a serious improvement over the rule-based model described in [21] and [76], which showed only 45.2 % on unknown words. Dots on accuracy graphs represent maximums on known (training set) and unknown (validation set) forms.

Having a closer look at some mistakes in Table 5, made by the character-to-character model in its best configuration (further referred as *char2char*), we can clearly see, that it learned to demutate forms (cf. the last two examples), but some inflection models are still unknown to it, which can be explained by the lack of training data. The model experiences most difficulties with compound verbs, which is not surprising.

As poor as the results may seem, they are not very different from those achieved by sequence-to-sequence models on analogous tasks. For example, the best results for the OCR post-

Figure 1: Character-to-character model accuracy

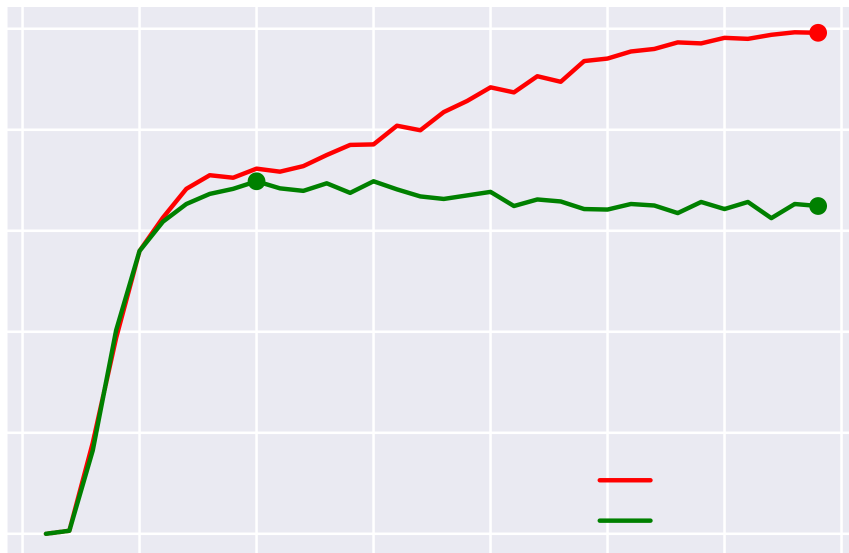

Table 6: Performance of different models on Early Irish data

Model	**Accuracy (unknown)**	**Accuracy (known)**
baseline	57.5 %	57.5 %
rule-based	45.2 %	71.6 %
char2char	64.9 %	99.2 %

correction and spelling correction tasks according to [59] fall between 62.75 % and 74.67 % on different datasets. The score is even lower for grapheme-to-phoneme task, 44.74 % – 72.23 % [59]. Lemmatisation scores described in the article are much higher, 94.22 % for German verbs and 94.08 % for Finnish verbs [59], but taking the inflectional diversity and abundant orthographic variation of Early Irish into account, this task is closer to spelling correction and grapheme-to-phoneme translation rather than to lemmatisation of any modern language. In any case, a character-level sequence-to-sequence model reached the accuracy score of 99.2 % for known words and 64.9 % for unknown words on a rather small corpus of 83,155 samples, which is a serious improvement over the rule-based model described in [21]. Table 6 shows the performance of different models on Early Irish data.

The model also meets the results of other systems working with historical data. Table 7 provides a summary of best accuracy scores achieved by Early Irish, Middle Dutch [37],

Table 7: Best accuracy scores on historical language data

Language	**Model**	**Unknown**	**Known**
Early Irish	character-level seq2seq	64.9 %	99.2 %
Middle Dutch	CNN + word embeddings	59.48 %	97.89 %
Latin	CRF	81.84 %	95.58 %
Old French	rule-based	?	60 %

Latin [47] and Old French [66] lemmatisers having different architectures are give in Table 7. Unfortunately, it is not possible to cite more results as there are no clear figures in other works concerning lemmatisation for ancient languages.

5 Conclusion

Although the task of lemmatisation for Early Irish data is quite challenging, there is a number of promising solutions. A character-level sequence-to-sequence model appears to be the best one for the moment, reaching the accuracy score of 99.2 % for known words and 64.9 % for unknown words on a rather small corpus of 83,155 samples. It outperforms both the baseline and the rule-based model and meets the results of other systems working with historical data.

Nevertheless, there is still much space for improvement and further research, and the first priority task that could help to ameliorate the performance is creating an open-source searchable corpus of Early Irish. It is also important to develop a detailed sensible grammatical notation to avoid such things as dropping out infixed pronouns when lemmatising verbal forms that persist in the DIL.

The results of the research, including working rule-based and seq2seq models and data, are available on GitHub.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Rami Al-Rfou, Guillaume Alain, Amjad Almahairi, Christof Angermueller, Dzmitry Bahdanau, Nicolas Ballas, Frédéric Bastien, Justin Bayer, Anatoly Belikov, Alexander Belopolsky, Yoshua Bengio, Arnaud Bergeron, James Bergstra, Valentin Bisson, Josh Bleecher Snyder, Nicolas Bouchard, Nicolas Boulanger-Lewandowski, Xavier Bouthillier, Alexandre de Brébisson, Olivier Breuleux, Pierre-Luc Carrier, Kyunghyun Cho, Jan Chorowski, Paul Christiano, Tim Cooijmans, Marc-Alexandre Côté, Myriam Côté, Aaron Courville, Yann N. Dauphin, Olivier Delalleau, Julien

Demouth, Guillaume Desjardins, Sander Dieleman, Laurent Dinh, Mélanie Ducoffe, Vincent Dumoulin, Samira Ebrahimi Kahou, Dumitru Erhan, Ziye Fan, Orhan Firat, Mathieu Germain, Xavier Glorot, Ian Goodfellow, Matt Graham, Caglar Gulcehre, Philippe Hamel, Iban Harlouchet, Jean-Philippe Heng, Balázs Hidasi, Sina Honari, Arjun Jain, Sébastien Jean, Kai Jia, Mikhail Korobov, Vivek Kulkarni, Alex Lamb, Pascal Lamblin, Eric Larsen, César Laurent, Sean Lee, Simon Lefrancois, Simon Lemieux, Nicholas Léonard, Zhouhan Lin, Jesse A. Livezey, Cory Lorenz, Jeremiah Lowin, Qianli Ma, Pierre-Antoine Manzagol, Olivier Mastropietro, Robert T. McGibbon, Roland Memisevic, Bart van Merriënboer, Vincent Michalski, Mehdi Mirza, Alberto Orlandi, Christopher Pal, Razvan Pascanu, Mohammad Pezeshki, Colin Raffel, Daniel Renshaw, Matthew Rocklin, Adriana Romero, Markus Roth, Peter Sadowski, John Salvatier, François Savard, Jan Schlüter, John Schulman, Gabriel Schwartz, Iulian Vlad Serban, Dmitriy Serdyuk, Samira Shabanian, Étienne Simon, Sigurd Spieckermann, S. Ramana Subramanyam, Jakub Sygnowski, Jérémie Tanguay, Gijs van Tulder, Joseph Turian, Sebastian Urban, Pascal Vincent, Francesco Visin, Harm de Vries, David Warde-Farley, Dustin J. Webb, Matthew Willson, Kelvin Xu, Lijun Xue, Li Yao, Saizheng Zhang, and Ying Zhang. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016.

[3] Mohammed Attia, Younes Samih, Khaled F Shaalan, and Josef Van Genabith. The floating Arabic dictionary: An automatic method for updating a lexical database through the detection and lemmatization of unknown words. In *COLING*, pages 83–96, 2012.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint*, 2014.

[5] Alistair Baron and Paul Rayson. VARD2: A tool for dealing with spelling variation in historical corpora. In *Postgraduate conference in corpus linguistics*, 2008.

[6] Alistair Baron and Paul Rayson. Automatic standartisation of texts containing spelling variation. *How much training data do you need*, 2009.

[7] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[8] Adam L Berger, Vincent J Della Pietra, and Stephen A Della Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71, 1996.

[9] Marcel Bollmann, Stefanie Dipper, Julia Krasselt, and Florian Petran. Manual and semi-automatic normalization of historical spelling-case studies from Early New High German. In *KONVENS*, pages 342–350, 2012.

[10] Lars Borin and Markus Forsberg. Something old, something new: A computational morphological description of Old Swedish. In *LREC 2008 workshop on language technology for cultural heritage data (LaTeCH 2008)*, pages 9–16, 2008.

[11] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint*, 2015.

[12] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint*, 2014.

[13] François Chollet et al. Keras. https://github.com/fchollet/keras, 2015.

[14] Grzegorz Chrupała. Simple data-driven context sensitive lemmatization. *Procesamiento del Lenguaje Natural*, 37:121–127, 2006.

[15] Grzegorz Chrupała. Normalizing tweets with edit scripts and recurrent neural embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 680–686. Citeseer, 2014.

[16] Grzegorz Chrupała, Georgiana Dinu, and Josef Van Genabith. Learning morphology with Morfette. 2008.

[17] Silvie Cinková and Jan Pomikálek. LEMPAS: A make-do lemmatizer for the Swedish PAROLE-corpus. *Prague Bull. Math. Linguistics*, 86:47–54, 2006.

[18] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[19] Walter Daelemans, Hendrik J Groenewald, and Gerhard B Van Huyssteen. Prototype-based active learning for lemmatization. 2009.

[20] Guy De Pauw and Gilles-Maurice De Schryver. Improving the computational morphological analysis of a Swahili corpus for lexicographic purposes. *Lexikos*, 18(1), 2008.

[21] Oksana Dereza. Building a dictionary-based lemmatizer for Old Irish. *Actes de la conférence conjointe JEP-TALN-RECITAL 2016, volume 6: CLTW*, pages 12–17, 2016.

[22] Elaine Uí Dhonnchadha. A two-level morphological analyser and generator for Irish using finite-state transducers. In *LREC*, 2002.

[23] Sander Dieleman, Jan Schlüter, Colin Raffel, Eben Olson, Søren Kaae Sønderby, Daniel Nouri, et al. Lasagne: First release., August 2015.

[24] Tarek El-Shishtawy and Fatma El-Ghannam. An accurate Arabic root-based lemmatizer for information retrieval purposes. *arXiv preprint*, 2012.

[25] Andrea Ernst-Gerlach and Norbert Fuhr. Retrieval in text collections with historic spelling using linguistic and spelling variants. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 333–341. ACM, 2007.

[26] Jesús Giménez and Lluis Marquez. SVMTool: A general POS tagger generator based on support vector machines. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*. Citeseer, 2004.

[27] Rafael Giusti, A Candido, Marcelo Muniz, Lívia Cucatto, and Sandra Aluísio. Automatic detection of spelling variation in historical corpus. In *Proceedings of the Corpus Linguistics Conference (CL)*, 2007.

[28] Péter Halácsy and V Trón. Benefits of deep NLP-based lemmatization for information retrieval. *CLEF (Working Notes)*, 2006.

[29] Iris Hendrickx and Rita Marquilhas. From old texts to modern spellings: An experiment in automatic normalisation. *JLCL*, 26(2):65–76, 2011.

[30] Xuedong D Huang, Yasuo Ariki, and Mervyn A Jack. *Hidden Markov models for speech recognition*, volume 2004. Edinburgh university press Edinburgh, 1990.

[31] Gérard Huet. Towards computational processing of Sanskrit. In *International Conference on Natural Language Processing (ICON)*. Citeseer, 2003.

[32] Anton Karl Ingason, Sigrún Helgadóttir, Hrafn Loftsson, and Eiríkur Rögnvaldsson. A mixed method lemmatization algorithm using a hierarchy of linguistic identities (HOLI). In *Advances in Natural Language Processing*, pages 205–216. Springer, 2008.

[33] Thorsten Joachims. *Learning to classify text using support vector machines: Methods, theory and algorithms*. Kluwer Academic Publishers, 2002.

[34] Kyle P. Johnson et al. Cltk: The classical language toolkit. https://github.com/cltk/cltk, 2014–2017.

[35] Jakub Kanis and Luděk Müller. Automatic lemmatizer construction with focus on OOV words lemmatization. In *International Conference on Text, Speech and "Dialogue"*, pages 132–139. Springer, 2005.

[36] Mike Kestemont, Walter Daelemans, and Guy De Pauw. Weigh your words—memory-based lemmatization for middle dutch. *Literary and Linguistic Computing*, 25(3):287–301, 2010.

[37] Mike Kestemont, Guy de Pauw, Renske van Nie, and Walter Daelemans. Lemmatization for variation-rich languages using deep learning. *Digital Scholarship in the Humanities*, page fqw034, 2016.

[38] Adam Kilgarriff, Michael Rundell, and Elaine Uí Dhonnchadha. Efficient corpus development for

lexicography: building the New Corpus for Ireland. *Language resources and evaluation*, 40(2):127–152, 2006.

[39] Julian Kupiec. Robust part-of-speech tagging using a hidden Markov model. *Computer Speech & Language*, 6(3):225–242, 1992.

[40] John Lafferty, Andrew McCallum, Fernando Pereira, et al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*, volume 1, pages 282–289, 2001.

[41] Elliott Lash. The parsed Old and Middle Irish corpus (POMIC). version 0.1. 2014.

[42] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[43] Dimitrios P Lyras, Kyriakos N Sgarbas, and Nikolaos D Fakotakis. Applying similarity measures for automatic lemmatization: A case study for Modern Greek and English. *International Journal on Artificial Intelligence Tools*, 17(05):1043–10 language = english,64, 2008.

[44] Nino Marinone. A project for Latin lexicography: 1. Automatic lemmatization and word-list. *Computers and the Humanities*, 24(5):417–420, 1990.

[45] Andrew McCallum, Dayne Freitag, and Fernando CN Pereira. Maximum entropy markov models for information extraction and segmentation. In *Icml*, volume 17, pages 591–598, 2000.

[46] Marieke Meelen and Barend Beekhuizen. PoS-tagging and chunking historical Welsh. In *Proceedings of the scottish celtic colloquium 2012*, 2013.

[47] Thomas Müller, Ryan Cotterell, Alexander M Fraser, and Hinrich Schütze. Joint lemmatization and morphological tagging with Lemming. In *EMNLP*, pages 2268–2274, 2015.

[48] David Packard. Computer-assisted morphological analysis of ancient Greek. 1973.

[49] Marco Carlo Passarotti. Development and perspectives of the Latin morphological analyser LEMLAT. *Linguistica computazionale*, 20(A):397–414, 2004.

[50] Snigdha Paul, Nisheeth Joshi, and Iti Mathur. Development of a Hindi lemmatizer. *arXiv preprint*, 2013.

[51] Praharshana Perera and René Witte. A self-learning context-aware lemmatizer for German. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 636–643. Association for Computational Linguistics, 2005.

[52] Thomas Pilz, Andrea Ernst-Gerlach, Sebastian Kempken, Paul Rayson, and Dawn Archer. The identification of spelling variants in english and german historical texts: manual or automatic? *Literary and Linguistic Computing*, 23(1):65–72, 2008.

[53] Michael Piotrowski. Natural language processing for historical texts. *Synthesis Lectures on Human Language Technologies*, 5(2):1–157, 2012.

[54] Joël Plisson, Nada Lavrac, Dunja Mladenic, et al. A rule based approach to word lemmatization. In *Proceedings C of the 7th International Multi-Conference Information Society IS 2004*, volume 1, pages 83–86. Citeseer, 2004.

[55] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[56] Martin Reynaert, Iris Hendrickx, and Rita Marquilhas. Historical spelling normalization. A comparison of two statistical methods: TICCL and VARD2. *on Annotation of Corpora for Research in the Humanities (ACRH-2)*, page 87, 2012.

[57] Raúl Rojas. Neural networks. a systematic introduction. *New York*, 1996.

[58] Yves Scherrer and Tomaž Erjavec. Modernizing historical Slovene words with character-based SMT. In *BSNLP 2013-4th Biennial Workshop on Balto-Slavic Natural Language Processing*, 2013.

[59] Carsten Schnober, Steffen Eger, Erik-Lân Do Dinh, and Iryna Gurevych. Still not there? Comparing traditional sequence-to-sequence models to encoder-decoder neural networks on monotone string translation tasks. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, page (to appear), December 2016.

[60] Djamé Seddah, Grzegorz Chrupała, Özlem Çetinoğlu, Josef Van Genabith, and Marie Candito.

Lemmatization and lexicalized statistical parsing of morphologically rich languages: the case of French. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 85–93. Association for Computational Linguistics, 2010.

[61] Ilya Segalovich. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *MLMTA*, pages 273–280. Citeseer, 2003.

[62] Fei Sha and Fernando Pereira. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 134–141. Association for Computational Linguistics, 2003.

[63] T. Shavrina and A. Sorokin. Modeling advanced lemmatization for Russian language using TnT-Russian morphological parser. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog"*, 2015.

[64] Richard Socher, Cliff C Lin, Chris Manning, and Andrew Y Ng. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 129–136, 2011.

[65] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, Christopher Potts, et al. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer, 2013.

[66] Gilles Souvay and Jean-Marie Pierrel. Lgerm lemmatisation des mots en moyen français. *Traitement Automatique des Langues*, 50(2):21, 2009.

[67] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[68] Gregory Toner, Grigory Bondarenko, Maxim Fomin, and Thomas Torma. An electronic dictionary of the Irish language. 2007.

[69] Kristina Toutanova and Colin Cherry. A global model for joint lemmatization and part-of-speech prediction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, volume 1, pages 486–494. Association for Computational Linguistics, 2009.

[70] Kristina Toutanova and Christopher D Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, pages 63–70. Association for Computational Linguistics, 2000.

[71] Aad Verboom. Towards a Sanskrit wordparser. *Literary and Linguistic Computing*, 3(1):40–44, 1988.

[72] Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv preprint*, 2015.

[73] Kaisheng Yao and Geoffrey Zweig. Sequence-to-sequence neural net models for grapheme-to-phoneme conversion. *arXiv preprint*, 2015.

[74] Т.А. Архангельский, Е.А. Мишина, and А.А. Пичхадзе. Система электронной грамматической разметки древнерусских и церковнославянских текстов и её использование в веб-ресурсах. In *Писменото наследство и информационните технологии. El'Manuscript–2014*, pages 102–104, 2014.

[75] Татьяна Сергеевна Гаврилова, Татьяна Александровна Шалганова, and Ольга Николаевна Ляшевская. К задаче автоматической лексико-грамматической разметки старорусского корпуса xv-xvii вв. *Вестник Православного Свято-Тихоновского гуманитарного университета. Серия 3: Филология*, 2(47), 2016.

[76] Оксана Дереза. Разработка программы-лемматизатора для древнеирландского языка. Курсовая работа, 2016.

[77] А.А. Зализняк. Грамматический словарь русского языка. Словоизменение. 1980.