Lemmatization for Ancient Languages: Rules or Neural Networks?

Oksana Dereza[1,2(✉)]

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

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. This task is often considered solved for most modern languages irregardless of their morphological type, but the situation is dramatically different for ancient languages. Rich inflectional system and high level of orthographic variation common to these languages together with lack of resources make lemmatising historical data a challenging task. It becomes more and more important as manuscripts are being extensively digitized now, but still remains poorly covered in literature. In this work, I compare a rule-based and a neural network based approach to lemmatisation in case of Early Irish (Old and Middle Irish are often described together as "Early Irish") data.

Keywords: Early Irish · Natural language processing
Under-resourced languages · Lemmatisation · Neural networks
Sequence-to-sequence learning

1 Introduction

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. Lemmatisation for ancient languages is still poorly covered in literature, although this task becomes more and more important as manuscripts are being extensively digitized.

D. Ustalov et al. (Eds.): AINL 2018, CCIS 930, pp. 35–47, 2018.
https://doi.org/10.1007/978-3-030-01204-5_4

There are two suitable approaches to this task that I will describe and compare in this article in regard to Early Irish data: a rule-based approach and character-based neural network models.

2 Related Works

The problem of NLP for historical languages first arose in the last quarter of the XXth century in regard to Ancient Greek [32], Sanskrit [20,47] and Latin [29,33] 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 [3,4], Old French [44], Old Swedish [6], Early New High German [5], historical Portuguese [17,19,39], historical Slovene [40], Middle Welsh [30] and Middle Dutch [24,25]. Historical data processing in general has been surveyed in a substantial monograph [37] and several articles [16,36]. Apart from corpus studies, there have emerged several open-source tools for historical language processing, such as a Classical Language Toolkit[1] [22], 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 [8,9,12,13,18,28,43,46].

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 [11], Icelandic [21], Czech [23], Slovene [38], German [35], Hindi [34], Arabic [1,15] 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 [32]. A Latin lemmatiser from the aforementioned Python library CLTK also uses stem and suffix lexicons. The best morphological analyser for Russian, Mystem, is based on Zalizniak grammatical dictionary [50]. 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 [42]. A morphological analyser of modern Irish used in New Corpus of Ireland is based on finite-state transducers and described in [14] and [26].

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

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

stems and suffixes and derives lemma classes itself from the working corpus [10]. 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 [31]. 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. For example, a system combining convolutional architecture that models orthography with distributional word embeddings that represent lexical context was successfully implemented by [25] 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 [27], 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] [45]. 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 the reader is to read as *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 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. To compile a lexicon for the rule-based lemmatiser and a training corpus for the neural network

[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...

lemmatiser, I crawled DIL's website, parsed HTML files and derived a set of rules to restore contractions and remove unnecessary markup. As a result, I got 83,155 unique form-lemma pairs. They 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 insufficient for getting extremely good results in lemmatisation for a language as morphologically complex and orthographically inconsistent as Early Irish.

Also, a test set was manually created to evaluate a rule-based system, because the DIL data cannot be used for evaluation in this case. It is described in detail in the next section.

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

- inconsistent use of length marks;
- in later texts mute vowels 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 2nd person plural object, and *-si* is an emphatic suffixed pronoun 3rd 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 thabair* 'does not give, bring'. Table 2 illustrates the variety of Early Irish verbal forms through the example of *do-beir*.

Table 2. Some forms of the verb 'do-beir'

Form	Deuterotonic	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'

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.

4 Rule-Based Approach

At first, I chose rule-based approach to lemma prediction over machine learning due to the scarcity of available data.

Morphophonological complexity of Early Irish compounded by the many non-transparent orthographic features makes traditional rule-based approach to lemmatisation with hard-coded lists of possible pseudo-suffixes and rules of their treatment less suitable for Early Irish than for other languages. A more reliable way for a start is building a full form lexicon where every word form corresponds to a lemma. I used the DIL described in the previous section for this purpose.

There was a series of experiments conducted on Early Irish prose texts that resulted into the following architecture of the rule-based lemmatiser. Every word in a text fed to the system is first demutated (i.e. the changes at the beginning of the word are eliminated) and then looked up in the dictionary. The lemmatiser returns a lemma for each known word and a demutated form for each unknown word by default; there is also an option to predict lemmas for unknown words with the help of Damerau-Levenshtein edit distance. For every unknown word, the program generates all possible strings on edit distance 1 and 2, checks them up in the dictionary and adds those that prove to be real words to the candidate list. Then the candidates are filtered by the first character: if the unknown word starts with a vowel, the candidate should also start with a vowel, and if the unknown word starts with a consonant, the candidate should start with the same consonant. Those parameters were chosen empirically as they yield the best results, i.e. the highest percentage of correctly predicted lemmas. Finally, the lemma of the candidate that has the highest probability is taken as a lemma for the unknown word.

In this work, I did not focus on word sense disambiguation, which means that if two or more different lemmas have identical forms, we cannot say for sure which lemma should be chosen for a particular instance of a homonymous form. The system provides two options for such cases: either return a list of all possible lemmas or choose the lemma with the highest probability. I should point out, that probability here is not a probability in a strict mathematical sense. Word form probability is formulated as a frequency count computed for each word in the test corpus, and lemma probability is the the sum of probabilities of forms belonging to a lemma.

Rule based lemmatiser was evaluated using accuracy score, which is a common metric for this task, on a set of manually annotated sentences, randomly chosen from Early Irish texts given in Table 3, which belong to different periods. The test set consists of 50 sentences, 840 tokens in total. It is worth mentioning, that the lemmatiser's lexicon contains mostly Old Irish forms with a small amount of Middle Irish ones and barely any Early Modern Irish ones. While Old Irish data helps to check how the system copes with unknown words' grammar, Middle and Early Modern Irish data is supposed to show its achievements with spelling variation. One also has to bear in mind, that the lemmatisers's performance is affected by form homonymy, which, given the absence of disambiguation, worsens the results.

Table 3. Early Irish texts used for creating a test set

Text	Period
Togail Bruidne Dá Derga	VII-IX centuries
Tochmarc Étaine	VIII-IX centuries
Fled Dúin na nGéd	XI-XII centuries
Lebor Gabála Érenn	XII century
Cath Finntrágha	XV century
Aided Muirchertaig Meic Erca	XIV-XV centuries
Buile Shuibhne	XVII-XVIII centuries

The system's performance is given in tables below; Table 4 compares the rule-based lemmatiser results with the baseline, defined as demutating a form, and Table 5 gives more detailed information.

Table 4. Rule-based model accuracy

Algorithm	Overall accuracy	Known words	Unknown words
Baseline	57.5%	57.5%	57.5%
Rule-based	65.7%	71.6%	45.2%

The system outperforms the baseline algorithm only by 8.2%, and these results are undoubtedly poor and not promising enough to continue the development of a rule-based system.

5 Neural Network Approach

The main problem Early Irish poses to machine learning methods is that its morphological complexity implies too many possible lemma classes, which, in addition to that, cannot always be reduced to a combination of a stem type and a suffix. Therefore some statistical models popular in sequence tagging that involves multi-class classification, such as HMM, MaxEnt, MEMM, SVM or CRF, are quite useless for this task. The best solution here seems to be turning from statistical machine learning to deep learning and using a sequence-to-sequence model, which allows going down to the character level. Basically, a sequence-to-sequence model is an ensemble of recurrent neural networks, or RNNs, that takes a sequence of a dynamic length as input and produces another sequence of a dynamic length. Sequence-to-sequence networks are used for a wide variety of tasks, such as grapheme-to-phoneme encoding [49], OCR post-processing, spelling correction, lemmatisation [41], machine translation [2,7] and even dialogue systems development [48]. Thus, if we reformulate the lemmatisation task as taking a sequence of characters (form) as input and generating

Table 5. Rule-based lemmatiser performance: details

Tokens	840
Known words	654
Unknown words	186
Lemmatised correctly	552
Predicted lemmas	157
Failed to predict	29
Predicted correctly	84
Predicted incorrectly	68
Disambiguation mistakes	73

another sequence of characters (lemma), we can forget about tens of verbal and nominal inflection classes, let alone spelling variation.

The data used in this experiment consists of 83,155 unique form-lemma pairs from the electronic edition of the DIL [45], shuffled and split into training, validation and test sets, the former two being 5,000 samples each. All experiments were run on a personal laptop with Intel Core i7 2,5 GHz processor and 12 Gb RAM, which took about 36 h each.

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 Fig. 1.

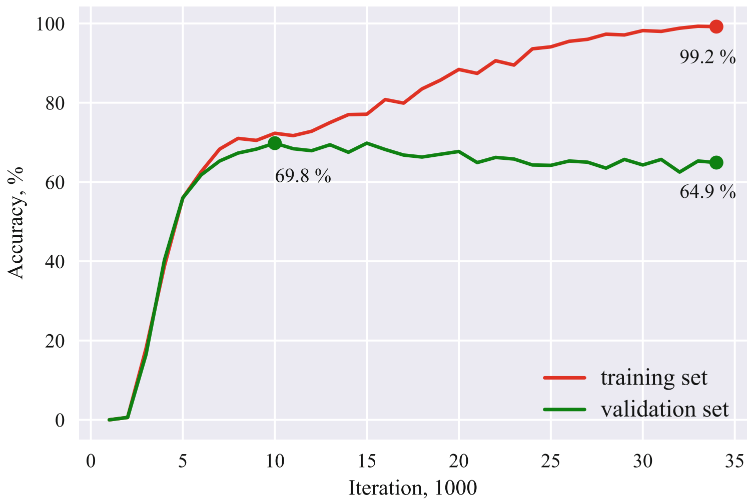

Fig. 1. Character-to-character model accuracy

These results are a serious improvement over the rule-based model, 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 some mistakes in Table 6, 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.

Table 6. 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

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-correction and spelling correction tasks according to [41] fall between 62.75% and 74.67% on different datasets. The score is even lower for grapheme-to-phoneme task, 44.74%–72.23% [41]. Lemmatisation scores described in the article are much higher, 94.22% for German verbs and 94.08% for Finnish verbs [41], 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 above. Table 7 shows the performance of different models on Early Irish data.

The model also meets the results of other systems working with historical data. Table 8 provides a summary of best accuracy scores achieved by Early Irish, Middle Dutch [25], Latin [31] and Old French [44] lemmatisers having different architectures. Unfortunately, it is not possible to cite more results as there are no clear figures in other works concerning lemmatisation for ancient languages.

Table 7. 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%

Table 8. 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%

6 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.

References

1. Attia, M., Samih, Y., Shaalan, K.F., Van Genabith, J.: The floating Arabic dictionary: an automatic method for updating a lexical database through the detection and lemmatization of unknown words. In: COLING, pp. 83–96 (2012)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint (2014). arXiv:1409.0473
3. Baron, A., Rayson, P.: VARD2: a tool for dealing with spelling variation in historical corpora. In: Postgraduate Conference in Corpus Linguistics (2008)
4. Baron, A., Rayson, P.: Automatic standartisation of texts containing spelling variation. How much training data do you need (2009)
5. Bollmann, M., Dipper, S., Krasselt, J., Petran, F.: Manual and semi-automatic normalization of historical spelling-case studies from Early New High German. In: KONVENS, pp. 342–350 (2012)

6. Borin, L., Forsberg, M.: Something old, something new: a computational morphological description of Old Swedish. In: LREC 2008 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008), pp. 9–16 (2008)
7. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint (2014). arXiv:1406.1078
8. Chrupała, G.: Simple data-driven context sensitive lemmatization. Procesamiento del Leng. Nat. **37**, 121–127 (2006)
9. Chrupała, G.: Normalizing tweets with edit scripts and recurrent neural embeddings. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. vol. 2, pp. 680–686. Citeseer (2014)
10. Chrupała, G., Dinu, G., Van Genabith, J.: Learning morphology with Morfette (2008)
11. Cinková, S., Pomikálek, J.: LEMPAS: a make-do lemmatizer for the Swedish PAROLE-corpus. Prague Bull. Math. Linguist. **86**, 47–54 (2006)
12. Daelemans, W., Groenewald, H.J., Van Huyssteen, G.B.: Prototype-based active learning for lemmatization (2009)
13. De Pauw, G., De Schryver, G.M.: Improving the computational morphological analysis of a Swahili corpus for lexicographic purposes. Lexikos **18**(1), 303–318 (2008)
14. Dhonnchadha, E.U.: A two-level morphological analyser and generator for Irish using finite-state transducers. In: LREC (2002)
15. El-Shishtawy, T., El-Ghannam, F.: An accurate Arabic root-based lemmatizer for information retrieval purposes. arXiv preprint (2012). arXiv:1203.3584
16. Ernst-Gerlach, A., Fuhr, N.: 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, pp. 333–341. ACM (2007)
17. Giusti, R., Candido, A., Muniz, M., Cucatto, L., Aluísio, S.: Automatic detection of spelling variation in historical corpus. In: Proceedings of the Corpus Linguistics Conference (CL) (2007)
18. Halácsy, P., Trón, V.: Benefits of deep NLP-based lemmatization for information retrieval. CLEF (Working Notes) (2006)
19. Hendrickx, I., Marquilhas, R.: From old texts to modern spellings: an experiment in automatic normalisation. JLCL **26**(2), 65–76 (2011)
20. Huet, G.: Towards computational processing of Sanskrit. In: International Conference on Natural Language Processing (ICON). Citeseer (2003)
21. Ingason, A.K., Helgadóttir, S., Loftsson, H., Rögnvaldsson, E.: A mixed method lemmatization algorithm using a hierarchy of linguistic identities (HOLI). In: Nordström, B., Ranta, A. (eds.) GoTAL 2008. LNCS (LNAI), vol. 5221, pp. 205–216. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85287-2_20
22. Johnson, K.P., et al.: CLTK: the classical language toolkit. https://github.com/cltk/cltk (2014–2017)
23. Kanis, J., Müller, L.: Automatic lemmatizer construction with focus on OOV words lemmatization. In: Matoušek, V., Mautner, P., Pavelka, T. (eds.) TSD 2005. LNCS (LNAI), vol. 3658, pp. 132–139. Springer, Heidelberg (2005). https://doi.org/10.1007/11551874_17
24. Kestemont, M., Daelemans, W., De Pauw, G.: Weigh your words–memory-based lemmatization for Middle Dutch. Lit. Linguist. Comput. **25**(3), 287–301 (2010)
25. Kestemont, M., de Pauw, G., van Nie, R., Daelemans, W.: Lemmatization for variation-rich languages using deep learning. Dig. Scholarsh. Humanit. **32**, 1–19 (2016)

26. Kilgarriff, A., Rundell, M., Dhonnchadha, E.U.: Efficient corpus development for lexicography: building the New Corpus for Ireland. Lang. Resour. Eval. **40**(2), 127–152 (2006)
27. Lash, E.: The parsed Old and Middle Irish corpus (POMIC). version 0.1 (2014)
28. Lyras, D.P., Sgarbas, K.N., Fakotakis, N.D.: Applying similarity measures for automatic lemmatization: a case study for Modern Greek and English. Int. J. Artif. Intell. Tools **17**(05), 1043–1064 (2008)
29. Marinone, N.: A project for Latin lexicography: 1. Automatic lemmatization and word-list. Comput. Humanit. **24**(5), 417–420 (1990)
30. Meelen, M., Beekhuizen, B.: PoS-tagging and chunking historical Welsh. In: Proceedings of the Scottish Celtic Colloquium 2012 (2013)
31. Müller, T., Cotterell, R., Fraser, A.M., Schütze, H.: Joint lemmatization and morphological tagging with Lemming. In: EMNLP, pp. 2268–2274 (2015)
32. Packard, D.: Computer-assisted morphological analysis of ancient Greek (1973)
33. Passarotti, M.C.: Development and perspectives of the Latin morphological analyser LEMLAT. Linguist. Comput. **20**(A), 397–414 (2004)
34. Paul, S., Joshi, N., Mathur, I.: Development of a Hindi lemmatizer. arXiv preprint (2013). arXiv:1305.6211
35. Perera, P., Witte, R.: A self-learning context-aware lemmatizer for German. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 636–643. Association for Computational Linguistics (2005)
36. Pilz, T., Ernst-Gerlach, A., Kempken, S., Rayson, P., Archer, D.: The identification of spelling variants in English and German historical texts: manual or automatic? Lit. Linguist. Comput. **23**(1), 65–72 (2008)
37. Piotrowski, M.: Natural language processing for historical texts. Synth. Lect. Hum. Lang. Technol. **5**(2), 1–157 (2012)
38. Plisson, J., Lavrac, N., Mladenic, D., et al.: A rule based approach to word lemmatization. In: Proceedings C of the 7th International Multi-Conference Information Society IS 2004, vol. 1, pp. 83–86. Citeseer (2004)
39. Reynaert, M., Hendrickx, I., Marquilhas, R.: Historical spelling normalization. A comparison of two statistical methods: TICCL and VARD2. In: Proceedings of Annotation of Corpora for Research in the Humanities (ACRH-2), p. 87 (2012)
40. Scherrer, Y., Erjavec, T.: Modernizing historical Slovene words with character-based SMT. In: BSNLP 2013–4th Biennial Workshop on Balto-Slavic Natural Language Processing (2013)
41. Schnober, C., Eger, S., Dinh, E.L.D., Gurevych, I.: 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), December 2016, to appear
42. Segalovich, I.: A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In: MLMTA, pp. 273–280. Citeseer (2003)
43. Shavrina, T., Sorokin, A.: Modeling advanced lemmatization for Russian language using TnT-Russian morphological parser. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" (2015)
44. Souvay, G., Pierrel, J.M.: Lemmatisation des mots en Moyen Français. Traitement Autom. Lang. **50**(2), 21 (2009)
45. Toner, G., Bondarenko, G., Fomin, M., Torma, T.: An electronic dictionary of the Irish language (2007)

46. Toutanova, K., Cherry, C.: 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, vol. 1, pp. 486–494. Association for Computational Linguistics (2009)
47. Verboom, A.: Towards a Sanskrit wordparser. Lit. Linguist. Comput. **3**(1), 40–44 (1988)
48. Vinyals, O., Le, Q.: A neural conversational model. arXiv preprint (2015). arXiv:1506.05869
49. Yao, K., Zweig, G.: Sequence-to-sequence neural net models for grapheme-to-phoneme conversion. arXiv preprint (2015). arXiv:1506.00196
50. Zaliznyak, A.A.: Grammatichesky slovar russkogo yazyka. Slovoizmenenie. Russian grammatical dictionary. Inflection (1980)