# The CLIN27 Shared Task

Tjong Kim Sang, Erik; Bollmann, Marcel; Boschker, Remko; Casacuberta, Francisco; Dietz, Feike; Dipper, Stefanie; Domingo, Miguel; van der Goot, Rob; van Koppen, Marjo; Ljubešic, Nikola

*Published in:*
Computational Linguistics in the Netherlands Journal

Computational Linguistics in the Netherlands Journal 7 (2017) 53-64 Submitted 05/2017; Published 12/2017

# The CLIN27 Shared Task: Translating Historical Text to Contemporary Language for Improving Automatic Linguistic Annotation

**Erik Tjong Kim Sang**[1] ERIKT@XS4ALL.NL
**Marcel Bollmann**[2] BOLLMANN@LINGUISTICS.RUB.DE
**Remko Boschker**[4] REMKO@INFORMATIETUIN.NL
**Francisco Casacuberta**[11] FCN@PRHLT.UPV.ES
**Feike Dietz**[10] F.M.DIETZ@UU.NL
**Stefanie Dipper**[2] DIPPER@LINGUISTICS.RUB.DE
**Miguel Domingo**[11] MIDOBAL@PRHLT.UPV.ES
**Rob van der Goot**[4] R.VAN.DER.GOOT@RUG.NL
**Marjo van Koppen**[10] J.M.VANKOPPEN@UU.NL
**Nikola Ljubešić**[7,12] NIKOLA.LJUBESIC@FFZG.HR
**Robert Östling**[8] ROBERT@LING.SU.SE
**Florian Petran**[2] PETRAN@LINGUISTICS.RUB.DE
**Eva Pettersson**[9] EVA.PETTERSSON@LINGFIL.UU.SE
**Yves Scherrer**[3,5] YVES.SCHERRER@UNIGE.CH
**Marijn Schraagen**[10] M.P.SCHRAAGEN@UU.NL
**Leen Sevens**[6] LEEN.SEVENS@KULEUVEN.BE
**Jörg Tiedemann**[5] JORG.TIEDEMANN@HELSINKI.FI
**Tom Vanallemeersch**[6] TALLEM@CCL.KULEUVEN.BE
**Kalliopi Zervanou**[10] K.A.ZERVANOU@UU.NL

[1] *Meertens Institute, Oudezijds Achterburgwal 185, 1012 DK Amsterdam, The Netherlands*

[2] *Ruhr-Universität Bochum, Universitätsstraße 150, 44801 Bochum, Germany*

[3] *University of Geneva, Rue De-Candolle 5, 1211 Genéve 4, Switzerland*

[4] *University of Groningen, Oude Kijk in 't Jatstraat 26, 9712 EK Groningen, The Netherlands*

[5] *University of Helsinki, Fabianinkatu 33, 00014 Helsinki, Finland*

[6] *KU Leuven, Blijde Inkomststraat 13, 3000 Leuven, Belgium*

[7] *Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia*

[8] *Department of Linguistics, Stockholm University, SE-106 91 Stockholm, Sweden*

[9] *University of Uppsala, Thunbergsvägen 3H, Uppsala, Sweden*

[10] *Utrecht University, Trans 10, 3512 JK Utrecht, The Netherlands*

[11] *Universitat Politècnica de València, Camí de Vera s/n, 46071 València (Spain)*

[12] *University of Zagreb, Ivana Lučića 3, 10000 Zagreb, Croatia*

## Abstract

The CLIN27 shared task evaluates the effect of translating historical text to modern text with the goal of improving the quality of the output of contemporary natural language processing tools applied to the text. We focus on improving part-of-speech tagging analysis of seventeenth-century Dutch. Eight teams took part in the shared task. The best results were obtained by teams employing character-based machine translation. The best system obtained an error reduction of 51% in comparison with the baseline of tagging unmodified text. This is close to the error reduction obtained by human translation (57%).

## 1. Introduction

In recent years, there has been a growing interest in using historical texts for linguistic research (Bennis 2011). However, many languages have considerably changed in the past few centuries (van der Sijs 2001) and as a result of this, computational tools which perform well on contemporary language may perform poorly on texts from several hundreds of years ago. This problem restricts the linguistic research that can be done with older texts.

In order to overcome this problem, we can take two actions. The first is to adapt the language processing tools and retrain them on appropriate language variants (Rem and van Halteren 2007). This requires a considerable amount of work in developing gold standard data for training the tools, a task which needs to be repeated for different time periods and for different tools. The second approach would be to translate the texts in the older variant of the language to a modern variant. Existing automatic annotation tools could then be applied to the new version of the text and the annotations could then be mapped to the original text. This approach would also require the development of gold standard training data for the translation system for each time period. However, translating texts is easier than making gold standard linguistic annotations and therefore the creation of this data would be cheaper, while it could benefit several tools.

In the CLIN27 shared task[1] we chose the second approach for adaptation of natural language processing tools to historical versions of a language. We restricted ourselves to the language Dutch as written in the seventeenth century and focused on improving the quality of syntactic word classes (part-of-speech tags). We would like to answer two questions. First, what techniques can be used for obtaining good translations from historical language to contemporary language? And second, how much does the quality of automatic annotations improve when we derive them from translated text in comparison with deriving them from the original historical text?

After an overview of related work, we will describe the data and the other resources of the shared task in Section 3. We will evaluate both translation quality and part-of-speech tag accuracy. The evaluation methods will be outlined in Section 4. Eight teams participated in the shared task and the results of their systems will be discussed in Section 5.

## 2. Related work

Growing interest in computational linguistic analysis of historical texts can be observed from recent workshops (at NoDaLiDa 2013 and 2017) and PhD theses (Rama 2015, Pettersson 2016). An important theme in the field is dealing with the spelling variation in the historical texts, caused by the absence of spelling standards (Bollmann et al. 2012). A commonly used tool for converting historical text to a standard spelling is VARD (Archer et al. 2015) which, among others, has been applied to English (Schneider et al. 2015) and Dutch (van Elburg and Wijckmans 2016). The system TiCCL has been designed specifically for spelling normalization of Dutch (Reynaert 2005). Tjong Kim Sang (2016) translated seventeenth-century Dutch text for improving part-of-speech tagging quality. Eisenstein (2013) discusses some of the problems of text normalization: it is not always clear what the standard to normalize to should be and the normalization step may change the meaning of a text. His work deals with social media text, a text type for which annotation tools suffer from similar problems as for historical text (Kaufmann and Kalita 2010, Han and Baldwin 2011, De Clercq et al. 2013).

Less work has been done on actually retraining tools for historical text. Hupkes (2014) used semi-supervised learning for tagging seventeenth-century Dutch. The system Adelheid was developed for tagging and lemmatizing Middle Dutch (Rem and van Halteren 2007). Yang and Eisenstein (Yang and Eisenstein 2016) replaced text tokens by vectors obtained from unsupervised learning to overcome the vocabulary differences between the contemporary training data for the tools and

1. Data and software used in the CLIN27 shared task are available at `http://ifarm.nl/clin2017st/`

the historical data. Their work is an example of domain adaptation which is also an interesting approach to our task (Jiang 2008, Plank 2011).

## 3. Data and resources

We would like to be able to apply tools developed for processing contemporary Dutch to texts that go back as far as 1100 AD (Brugman et al. 2016). For this purpose we want to translate historical Dutch to modern Dutch. However, there is not a single variant of historical Dutch, but several variants, like Old Dutch, Middle Dutch and Early New Dutch, which overlap. We chose to work with texts of the seventeenth century, firstly because this period is of prime interest in Dutch history studies (the so-called Dutch Golden Age) and secondly because it is situated approximately in the middle of the focus eras of two important part-of-speech taggers for Dutch: Frog: twenty-first century (Van den Bosch et al. 2007) and Adelheid: fourteenth century (Rem and van Halteren 2007).

The seventeenth century has as an additional benefit that it is included in an existing parallel corpus: the Statenvertaling Bible (about 925,000 tokens) which is available in digital format in editions from 1637, 1657, 1888 and 2010 (van der Sijs 2008, Beelen and van der Sijs 2014, Theologencommissie 1999, Stichting Herziene Statenvertaling 2010). For our purposes, the Statenvertaling has as an advantage over other modernized historical works that care has been taken to keep the new editions as close as possible to the original one from 1637. Most of the tokens of the 1637 edition can be mapped to counterparts in the 1657 and 1888 editions, which makes it easy to align the different versions of the text and extract useful parallel lexicons for the translation task. The relation between the 2010 edition and the original is more free: there is a perfect one-to-one sentence mapping between the two, but between linked sentences there are several differences in token counts and token orders.

Here is an example sentence from the four Bible editions with modified tokens marked in red:

> **1637**: *De Aerde nu was woest ende ledich , ende duysternisse was op den afgront : ende de Geest Godts sweefde op de Wateren .*
>
> **1657**: *De Aerde nu was woest ende ledigh , ende duysternisse was op den afgrondt : ende de Geest Godts sweefde op de Wateren.*
>
> **1888**: *De aarde nu was woest en ledig , en duisternis was op den afgrond ; en de Geest Gods zweefde op de wateren .*
>
> **2010**: *De aarde nu was woest en leeg , en duisternis lag over de watervloed ; en de Geest van God zweefde boven het water .*
>
> **English**: *The earth now was barren and empty , and darkness was upon the abyss : and the spirit of God glided over the waters .*

We offered the Statenvertaling texts as a basic training resource to the shared task participants. However, we did not want to evaluate the participating systems on text from a single domain (in this case: the Bible). Since there was no other suitable parallel text available, we decided to develop some ourselves by selecting seventeenth century Dutch texts and manually translating them to contemporary Dutch. As Eisenstein (2013) explained, such a text normalization step can be performed in different ways. For example, we could focus on making the text understandable for readers, replacing all historic words by modern versions and rearranging word order wherever necessary. But in that case the relation between the tokens in the translation and the ones in the original text would be complex and it would be difficult to map the annotations of the modern text back to the historical text. Another approach would be to replace only the historic tokens that are unknown to the tagger. But this approach would not improve the readability of the texts a lot and it would make the translation step tool-dependent.

| Task | Gold PoS | Year | Size | Author |
|---|---|---|---|---|
| train | - | 1637 | 1,107,567 | team of translators |
| train | - | 1888 | 1,087,536 | Jongbloed and others |
| develop | - | 1698 | 1,236 | Steven Blankaart |
| test | - | 1607 | 1,265 | P.C. Hooft |
| test | + | 1616 | 1,218 | G.A. Bredero |
| test | - | 1626 | 1,166 | Isaac Beeckman |
| test | - | 1636 | 1,226 | Hugo de Groot |
| test | - | 1646 | 1,187 | Willem Frederik van Nassau |
| test | - | 1656 | 1,153 | Jan van Riebeeck |
| test | - | 1668 | 1,179 | Hendrick Hamel |
| test | - | 1678 | 1,140 | Antoni Leeuwenhoek |
| test | - | 1686 | 1,331 | Govert Bidloo |
| test | - | 1692 | 1,207 | Constantijn Huygens jr. |
| extra | - | 1602 | 1,149 | Anthony Duyck |
| extra | - | 1621 | 1,268 | Noordsche Compagnie |
| extra | - | 1641 | 1,212 | Jacob Mahieusen |
| extra | - | 1662 | 1,103 | Staten Zeeland |
| extra | - | 1682 | 1,187 | Cornelis de Bruyn |

Table 1: Texts from the Digital Library for Dutch Literature[2] from which snippets of 1,100-1,400 tokens have been used in the shared task. All texts have been translated to contemporary Dutch. Gold standard part-of-speech tags were available for one test text.

While it would have been nice if the translations improved the readability of historical texts, we decided not to aim for this and focus on creating translations that benefited the part-of-speech tagger. In the additional data sets for the shared task, we only replaced tokens that did not occur in a standard lexicon with the required base syntactic category. As standard lexicon, we chose the Van Dale dictionary for Dutch (den Boon and Hendrickx 2015). This is generally accepted to be the standard dictionary for the language Dutch. By using this resource we can generate tool-independent translations.

Here is an example of a sentence in historical Dutch and its translation to modern Dutch. Like in the previous examples, tokens that were changed by the translation process have been marked in red:

> **original**: *'t Geslacht , de geboort , plaets , tydt , leven , ende wercken Van Karel van Mander , Schilder , en Poeet , Mitsgaders Zyn overlyden , ende begraeffenis .*
>
> **translation**: *'t Geslacht , de geboorte , plaats , tijd , leven , ende werken van Karel van Mander , schilder , en poëet , mitsgaders zijn overlijden , ende begrafenis .*
>
> **English**: *The gender , the birth , place , time, life , and work of Karel van Mander , painter , and poet , as well as his death , and funeral .*

For this sentence, only small orthographic changes were necessary but even such small changes can be very useful for the tagger. Note that certain archaic words like *ende* (*and*) and *mitsgaders* (*as well as*) have remained in the translation because they were listed in the Van Dale dictionary. Names pose a challenge for the translation process. We have left most names unchanged in the translation with the exception of a few frequent geographic names. We allow one-to-many and many-to-one

2. http://dbnl.nl

| CB | Lev | Mos | +da | rew | MG | cas | NN | com | IN | Team |
|---|---|---|---|---|---|---|---|---|---|---|
| - | - | - | - | + | + | - | - | - | + | Amsterdam |
| + | - | - | - | - | + | - | + | - | - | Bochum |
| - | + | - | + | - | - | + | - | - | - | Groningen |
| + | + | + | - | - | - | + | + | + | + | Helsinki/Uppsala |
| - | + | + | + | + | - | + | - | - | - | Leuven |
| + | + | + | + | + | + | + | - | - | - | Ljubljana/Zagreb/Geneva |
| - | - | + | - | - | - | + | - | - | - | Utrecht |
| - | - | + | + | - | + | + | - | - | - | Valencia |

Table 2: Features of the eight systems that participated in the CLIN27 shared task: CB: character-based translation methods, Lev: Levenshtein distance, Mos: Moses machine translation system, +da: used external data, rew: string rewrite rules, MG: MGIZA system, cas: case-sensitive processing, NN: neural network, com: system combination, IN: used INT lexicon

token alignments with the provision that single tokens may only be linked to sequences of successive tokens.

All the additional data originate from the Digital Library for Dutch Literature. In order to achieve a balanced corpus, we have selected small snippets of text (1,100-1,400 tokens ) from different authors and from different time periods: one text as development data, ten texts as test data and five texts for usage in future work, see Table 1. The five extra data sets were created after the shared task and were not available to the shared task participants. All texts have been translated to contemporary Dutch by a single annotator. For one of the test texts, Bredero 1616, gold standard base part-of-speech tags were created.

Apart from the Bible texts and the texts from DBNL, the participants also had access to a web service which offered looking up lemmas for contemporary and historical Dutch words: the INT lexicon service (INT 2015). While the service has an excellent coverage of historic Dutch words, using it for machine translation presented two challenges. First, the service returned modern lemmas rather than modern words, which means that morphological information can be lost. And second, the service may return several lemmas for one word. without any corpus frequency information. In such cases, other resources are necessary to make a reasonable choice among the alternative lemmas.

## 4. Evaluation

We evaluated two aspects of the translation task: the quality of the translated texts in comparison with human translations and the accuracy of a standard tagger applied to the translated texts in comparison with a gold standard. For evaluating the quality of the translations, we used the standard for such evaluations: BLEU (Papineni et al. 2002), which is based on the number of corresponding token n-grams in a translated text and the gold standard. We took the common settings for BLEU, as described in Papineni et al. (2002) and used in the translation system Moses (Koehn 2017): compare all token n-grams of lengths one to four in the same sentence. We always used a single text as the gold standard translation.

Although BLEU is the standard method for evaluation of machine translation tasks, it is not without problems. Already in 2006, Callison-Burch et al. (2006) showed that the correlation between BLEU evaluation and human evaluation can be poor. Therefore we performed an additional evaluation of the output of the shared task systems based on part-of-speech tagging accuracy. The Bredero 1616 text of each participant team has been processed with the same state-of-the-art tagger (Frog, see Van den Bosch et al. (2007)). The base tags in the results have been compared with a gold standard and the systems have been ranked by tag accuracy. Both for the accuracy scores

| Ranks | BLEU score | Team |
|---|---|---|
| 1 | 0.610±0.017 | Ljubljana/Zagreb/Geneva |
| 1 | 0.607±0.018 | Bochum |
| 1 | 0.599±0.017 | Helsinki/Uppsala |
| 4 | 0.568±0.019 | Amsterdam |
| 5 | 0.538±0.016 | Leuven |
| 6 | 0.469±0.017 | Utrecht |
| 6/7 | 0.451±0.017 | Groningen |
| 7 | 0.430±0.017 | Valencia |
| | 0.331±0.016 | baseline (unmodified text) |

Table 3: Results of the CLIN27 shared task measured by BLEU scores. The numbers ($e_i$) to the right of the scores are estimations of significance intervals ($p<0.05$). Scores are significantly different if they are more than $\sqrt{e_1^2 + e_2^2}$ apart. There is no significant difference between systems that share a rank number.

and the BLEU scores, we have estimated confidence intervals ($p<0.05$) with bootstrap resampling (Noreen 1989, Yeh 2000).

As a baseline for the translation quality evaluation, we compute the BLEU score of the historical test text in comparison with the manually translated text. The accuracy of the part of speech tags assigned to this historical text, serves as a baseline in the part of speech tag quality evaluation. The accuracy of the tags assigned to the manually translated text is used as a ceiling score in the tag evaluation. The performance of the shared task systems is not expected to become higher than this score that is based on human translation.

## 5. Results

Eight teams participated in the CLIN27 shared task on translating historical text (see Table 2). They used a variety of methods for performing the task. Most teams approached it as a machine translation task and the popular machine translation system Moses (Koehn 2017) and the associated MERT training algorithm (Och 2003) and MGIZA alignment were used by all but two of the teams (for example by Domingo et al. (2017)). The best three teams by BLEU score all explicitly mentioned employing character-based translation methods (for example Bollmann et al. (2017)). The Levenshtein distance was used by half of the systems. Two teams employed neural networks in one way or another.

With respect to BLEU score, the best performance was achieved by the team from Ljubljana/Zagreb/Geneva (Ljubešić and Scherrer 2017) with Bochum (Bollmann et al. 2017) and Helsinki-/Uppsala finishing second and third (see Table 3). The scores of the top-3 are not significantly different. All systems performed a lot better than the baseline system which returned the original historical text without any change.

The best performing system by Ljubešić and Scherrer (2017), consisted of a character-level statistical machine learning system (Moses) trained on the parallel Bible texts. Only aligned sentence pairs with a Levenshtein score of 0.7 or larger were used. Three different language models were used: one was learned from the Bibles from the shared task while the other two were learned from external data: Dutch subtitles (Lison and Tiedemann 2016) and the Dutch EUbookshop corpus (Skadiņš et al. 2014) The weights of the system were determined with the MERT algorithm (Och 2003). The extra parallel train file, Blankaart 1698, was used as development data. The system was trained

| Ranks | Accuracy | OOV | Team |
|---|---|---|---|
| | 0.875±0.026 | 0.093 | ceiling (tagged manual translation) |
| 1 | 0.856±0.029 | 0.148 | Helsinki/Uppsala |
| 1/2 | 0.842±0.022 | 0.131 | Ljubljana/Zagreb/Geneva |
| 1/2 | 0.836±0.027 | 0.129 | Amsterdam |
| 1/2/4 | 0.825±0.025 | 0.133 | Leuven |
| 1/2/4 | 0.824±0.030 | 0.085 | Bochum |
| 2/4 | 0.812±0.030 | 0.223 | Utrecht |
| 4/7 | 0.793±0.024 | 0.019 | Groningen |
| 7 | 0.759±0.030 | 0.282 | Valencia |
| | 0.709±0.030 | 0.407 | baseline (tagged unmodified text) |

Table 4: Results of the CLIN27 shared task measured by part-of-speech accuracy. The numbers to the right of the accuracy scores are estimations of significance intervals ($p<0.05$). The OOV column contains the rate of out-of-vocabulary words in the texts according to the tagger. There is no significant difference between systems that share a rank number.

on entire sentences and consequently normalizes an entire sentence at the time which could be an advantage. Previously this system has been used successfully for normalizing Slovene (Ljubešić et al. 2016) and normalizing Swiss German (Scherrer and Ljubešić 2016).

Measured by part-of-speech accuracy, the team from Helsinki/Uppsala performed best, followed by Ljubljana/Zagreb/Geneva and Amsterdam (see Table 4). Here the top-5 scores are not significantly different. Again all systems outperformed the baseline which was measured by tagging the unchanged seventeenth century text with the contemporary part-of-speech tagger. There was no significant difference between the score of the two best systems (error reduction compared with the baseline: 51% and 46%) and the base tag accuracy achieved by applying the tagger to a human translation of the test text (error reduction: 57%). The rate of out-of-vocabulary words (OOV) for the tagger proved to be a poor predictor of the performance of the systems: the system with the lowest OOV score did not achieve one of the best tagging accuracies (Table 4).

The team behind the best system, Östling, Petterson and Tiedemann (2017), evaluated three different approaches. The first was a convolutional neural network with character-level alignments trained on the Bibles. The second approach was a spelling normalization method based on the INT lexicon using Bible token frequencies to resolve ties with a back-off to Levenshtein measure search in the Bible texts for near neighbors. The third method was a phrase-based and character-based machine translation system trained on the Bible text and tuned on the Blankaart 1698 text with among others the MERT algorithm (Och 2003). Additionally, three different combinations of the systems were created by generating sentences which were closest by Levenshtein distance to the sentences of the systems participating in the combination. The combination of all three approaches performed best with respect to BLEU score. However, the systems using the neural network did not provide a word alignment so the part-of-speech tagging accuracies have only been computed for the other two methods. The statistical machine translation approach performed the best of the two as well as the best of all entries in the shared task.

All participating systems clearly improved on the baselines set for the shared task. But such improvements could be the result of trivial actions, like the normalization of a few very frequent closed class words. In order to check if this was the case, we performed an extensive analysis of the output of the Helsinki system of the Bredero 1616 text. We counted the number of times that the system replaced one token by another and the number of times these changes caused the base part-of-speech tag of a token to become different: 495 tokens were changed and this caused 305 tags to change. We found no frequent pairs dominating either of the two formats: the two most frequent token changes covered only 3% of the token changes each, while the two most frequent tag changes

| Word changes | | | Part-of-speech changes | | |
|---|---|---|---|---|---|
| 16x | +1.000 | ende→en | 37x | +0.595 | SPEC→N |
| 15x | +0.600 | 't→het | 36x | +0.556 | N→WW |
| 10x | +1.000 | so→zo | 29x | +0.414 | N→ADJ |
| 10x | +1.000 | ick→ik | 22x | +0.818 | N→VNW |
| 7x | +1.000 | sal→zal | 17x | +1.000 | SPEC→VG |
| 6x | +1.000 | zyn→zijn | 17x | +0.118 | SPEC→ADJ |
| 6x | +1.000 | soo→zo | 15x | +0.933 | SPEC→VNW |
| 6x | +1.000 | hy→hij | 15x | +0.133 | WW→N |
| 6x | +0.000 | daer→daar | 14x | +1.000 | SPEC→BW |
| 413x | +0.278 | ...others... | 103x | +0.621 | ...others... |

Table 5: Analysis of the output of the Helsinki SMT system on the Bredero 1616 text. We found that the system made 495 token changes (left) which in turn caused 305 base tag changes (right). Neither of the two formats were dominated by a single or a few frequent corrections. The system contributed to a wide variety of improvements.

dealt with 12% of the tag changes (see Table 5). We can conclude that the system makes a wide variety of changes and that its contributions seem to be valuable.

## 6. Concluding remarks

We presented the CLIN27 shared task: automatically translating historical text to modern text for improving the quality of part-of-speech tagging applied to the text. Eight teams participated in the task. The participants had access to parallel bible texts, a small parallel development text and an online lexicon. Some used additional historical text material for language modeling. The system of Ljubljana/Zagreb/Geneva performed best in the evaluation of translation quality (BLEU) and the system from Helsinki/Uppsala reached the highest score with respect to part-of-speech tagging quality. We found that the part-of-speech tag quality of the best system was close to the one obtained by human translation of the text but we should note that our test text was small.

The shared task has provided answers to the research questions put forward in the introduction of this paper. With respect to the techniques which could be used for obtaining good translations from historical text: we found several. Character-based translation deserves a special mention, as it was used by the best-performing systems. This confirms earlier results on similar tasks, for example De Clercq (2013). The machine translation system Moses was used by several shared task participants, just like Levenshtein-based methods for selecting the most appropriate translations.

We observed considerable performance gains, both with respect to text quality and attainable part-of-speech accuracy. We only had one version of the manually created test sets. It would have been interesting to have more versions available so that we could compare the systems with the performance that humans obtain for this task. Such a comparison was possible for tag accuracy and we found that the scores of the two best systems were not significantly worse than human performance. However, their performance is still well below the state-of-the-art for tagging contemporary language.

After the shared task, we see several directions for follow-up work. The systems participating in the shared task came close to the human performance for this approach, but the part-of-speech accuracies obtained were still more than ten percentage points lower than the state of the art for tagging contemporary text. It would be interesting to find out if the ceiling score is a hard boundary for these types of texts, or if higher accuracies are possible. An extensive error analysis of the system output could provide useful insights.

Another open question is whether this approach of translation or text normalization will benefit other natural language processing tools. An obvious one to evaluate next would be parsing, but it would also be interesting to look at the effects of translation to contemporary text on lemmatizing, morphological analysis, named entity recognition and semantic analysis.

Finally, the approach of this shared task could be used for dealing with texts from other periods and from other languages. A challenge is the selection of the time frames. Language changes gradually, so various selections of the past centuries could be made. Some could even be overlapping. We hope that this shared task can be an example on how to deal with part-of-speech tagger adaptions for processing other time periods and languages.

## Acknowledgments

We would like to thank three anonymous reviewers for valuable feedback on an earlier version of this paper. Erik Tjong Kim Sang was funded by the NWO-sponsored project Nederlab. Marcel Bollmann, Stefanie Dipper, and Florian Petran were supported by Deutsche Forschungsgemeinschaft (DFG), Grant DI 1558/4. Francisco Casacuberta and Miguel Domingo were funded by the Ministerio de Economía y Competitividad (MINECO) under project CoMUN-HaT (grant agreement TIN2015-70924-C2-1-R), and Generalitat Valenciana under project ALMAMATER (grant agreement PROMETEOII/2014/030). Nikola Ljubešić was supported by the Slovene research agency grant J6-6842 (project JANES). Marjo van Koppen, Feike Dietz, Marijn Schraagen and Kalliopi Zervanou were funded by the NWO-VC project Language Dynamics in the Dutch Golden Age, project number 360-78-020.

## References

Archer, Dawn, Merja Kyto, Alistair Baron, and Paul Rayson (2015), Guidelines for normalising Early Modern English corpora: Decisions and justifications, *ICAME Journal*. DOI: 10.1515/icame-2015-0001.

Beelen, Hans and Nicoline van der Sijs, editors (2014), *Biblia, dat is De gantsche H. Schrifture (Statenvertaling 1657)*, DBNL: Digitale Bibliotheek voor de Nederlandse Letteren. http://dbnl.nl/tekst/_sta001stat02_01/ (retrieved 10 May 2017).

Bennis, Hans (2011), *Nederlab: NWO Large application*, NWO.

Bollmann, Marcel, Joachim Bingel, and Anders Søgaard (2017), Learning attention for historical text normalization by learning to pronounce, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL2017)*, Vancouver, Canada.

Bollmann, Marcel, Stefanie Dipper, Julia Krasselt, and Florian Petran (2012), Manual and Semi-automatic Normalization of Historical Spelling – Case Studies from Early New High German, *Proceedings of the LThist workshop at KONVENS 2012*, Vienna, Austria.

Brugman, Hennie, Martin Reynaert, Nicoline van der Sijs, René van Stipriaan, Erik Tjong Kim Sang, and Antal van den Bosch (2016), Nederlab: Towards a Single Portal and Research Environment for Diachronic Dutch Text Corpora, *Proceedings of LREC 2016*, ELRA, Portoroz, Slovenia.

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn (2006), Re-evaluating the Role of BLEU in Machine Translation Research, *Proceedings of the 11th Conference of the European Chapter of Computational Linguistics (EACL 2006)*, ACL, Trento, Italy.

De Clercq, Orphée, Sarah Schulz, Bart Desmet, Els Lefever, and Véronique Hoste (2013), Normalization of Dutch User-Generated Content, *9th International conference on Recent Advances in*

*Natural Language Processing (RANLP 2013) (Hissar, Bulgaria)*, INCOMA (Shoumen, Bulgaria), pp. 179–188.

den Boon, C.A. and Ruud Hendrickx (2015), *Van Dale Groot woordenboek van de Nederlandse taal*, Van Dale.

Domingo, Miguel, Mara Chinea-Rios, and Francisco Casacuberta (2017), Historical documents modernization, *Proceedings of the Annual Conference of the European Association for Machine Translation*, Prague, Czech Republic.

Eisenstein, Jacob (2013), What to do about bad language on the internet, *Proceedings of NAACL-HLT 2013*, Association for Computational Linguistics, Atlanta, Georgia, pp. 359–369.

Han, Bo and Timothy Baldwin (2011), Lexical normalisation of short text messages: Makn sens a# twitter, *Proceedings of ACL HLT 2011*, Association for Computational Linguistics, Portland, OR, pp. 368–378.

Hupkes, Dieuwke (2014), Semi-supervised training of part of speech taggers for historical Dutch using modern Dutch syntactic priors. unpublished manuscript.

INT (2015), Lexicon Service. Instituut voor Nederlandse Taal, http://sk.taalbanknederlands.inl.nl/LexiconService/ Retrieved 7 May 2017.

Jiang, Jing (2008), A literature survey on domain adaptation of statistical classifiers, *http://sifaka.cs.uiuc.edu/jiang4/domain_adaptation/survey/da_survey.pdf Retrieved 13 May 2016*.

Kaufmann, Max and Jugal Kalita (2010), Syntactic normalization of twitter messages, *International conference on natural language processing (ICON)*, Kharagpur, India.

Koehn, Philipp (2017), *MOSES - Statistical Machine Translation System - User Manual and Code Guide*, University of Edinburgh.

Lison, Pierre and Jörg Tiedemann (2016), Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles, *in* Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, European Language Resources Association (ELRA), Paris, France.

Ljubešić, Nikola and Yves Scherrer (2017), cSMTiser submission for the CLIN2017 shared task, *Abstract submitted to Computational Linguistics in the Netherlands 27 (CLIN27)*, Leuven, Belgium.

Ljubešić, Nikola, Katja Zupan, Darja Fišer, and Tomaž Erjavec (2016), Normalising Slovene data: historical texts vs. user-generated content, *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, Bochumer Linguistische Arbetsberichte, Bochum, Germany.

Noreen, Eric W. (1989), *Computer-Intensive Methods for Testing Hypotheses*, John Wiley & Sons.

Och, Franz Josef (2003), Minimum error rate training in statistical machine translation, *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL'03)*, ACL, Sapporo, Japan, pp. 160–167.

Östling, Robert, Eva Pettersson, and Jörg Tiedemann (2017), The CLIN2017 shared task, *Abstract submitted to Computational Linguistics in the Netherl ands 27 (CLIN27)*, Leuven, Belgium.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002), BLEU: a method for Automatic Evaluation of Machine Translation, *Proceedings of ACL 2002*, Association for Computational Linguistics, Philadelphia PA, pp. 311–318.

Pettersson, Eva (2016), *Spelling Normalisation and Linguistic Analysis of Historical Text for Information Extraction*, PhD thesis, Studia Liguistica Upsaliensia 17, Uppsala: Acta Universitatis Upsaliensis.

Plank, Barbara (2011), *Domain Adaptation for Parsing*, PhD thesis, University of Groningen, The Netherlands.

Rama, Tarak (2015), *Studies in computationalhistorical linguistics – Models and analyses*, PhD thesis, Data linguistica 27, University of Gothenburg, Sweden.

Rem, M. and H. van Halteren (2007), *Tagging and Lemmatization Manual for the corpus van Reenen-Mulder and the Adelheid 1.0 Tagger-Lemmatizer*, Radboud University Nijmegen.

Reynaert, Martin (2005), *Text-Induced Spelling Correction*, PhD Thesis, Tilburg University.

Scherrer, Yves and Nikola Ljubešić (2016), Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation, *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, Bochumer Linguistische Arbetsberichte, Bochum, Germany.

Schneider, Gerold, Hans Martin Lehmann, and Peter Schneider (2015), Parsing Early and Late Modern English corpora, *Digital Scholarship in the Humanities*.

Skadiņš, Raivis, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne (2014), Billions of parallel words for free: Building and using the eu bookshop corpus, *in* Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, European Language Resources Association (ELRA), Reykjavik, Iceland.

Stichting Herziene Statenvertaling (2010), *Herziene Statenvertaling*, Stichting Herziene Statenvertaling. https://herzienestatenvertaling.nl/ (retrieved 10 May 2017).

Theologencommissie, editor (1999), *Statenvertaling Jongbloededitie 1888*, Statenvertaling.net. http://www.statenvertaling.net/ (retrieved 10 May 2017).

Tjong Kim Sang, Erik (2016), Improving Part-of-Speech Tagging of Historical Text by First Translating to Modern Text, *in* Bozic and Mendel-Gleason and Debruyne and O'Sullivan, editor, *2nd IFIP International Workshop on Computational History and Data-Driven Humanities*, Springer Verlag.

Van den Bosch, A., G.J. Busser, W. Daelemans, and S. Canisius (2007), An efficient memory-based morphosyntactic tagger and parser for Dutch, *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, Leuven, Belgium, pp. 99–114.

van der Sijs, Nicoline (2001), *Chronologisch woordenboek: De ouderdom en herkomst van onze woorden en betekenissen*, Veen, Amsterdam/Antwerpen.

van der Sijs, Nicoline, editor (2008), *Biblia, dat is De gantsche H. Schrifture (Statenvertaling 1637)*, DBNL: Digitale Bibliotheek voor de Nederlandse Letteren. http://dbnl.nl/tekst/_sta001stat01_01/ (retrieved 10 May 2017).

van Elburg, Wouter and Tessa Wijckmans (2016), Project COLEM for CREATE (University of Amsterdam)Adapting NPL-Tools for Creating an Orthographic Layer for Early Modern Dutch Texts, *Digital Humanities 2016: Conference Abstracts*, Jagiellonian University & Pedagogical University, Kraków, p. 906.

Yang, Yi and Jacob Eisenstein (2016), Part-of-speech tagging for historical english, *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, ACL, San Diego, CA.

Yeh, Alexander (2000), More accurate tests for the statistical significance of result differences, *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Saarbruecken, Germany, pp. 947–953.