Evaluating Inter-Annotator Agreement on Historical Spelling Normalization

Marcel Bollmann **Stefanie Dipper** **Florian Petran**
Ruhr-Universität Bochum
Department of Linguistics
{bollmann,dipper,petran}@linguistics.rub.de

Abstract

This paper deals with means of evaluating inter-annotator agreement for a normalization task. This task differs from common annotation tasks in two important aspects: (i) the class of labels (the normalized wordforms) is open, and (ii) annotations can match to different degrees. We propose a new method to measure inter-annotator agreement for the normalization task. It integrates common chance-corrected agreement measures, such as Fleiss's κ or Krippendorff's α. The novelty of our proposed method lies in the way the annotated word forms are treated. First, they are evaluated character-wise; second, certain characters are mapped to more general categories.

1 Introduction

In recent years, and in particular in the context of digital humanities, historical language data has been gaining increasing significance. The focus is on providing easy access to the information contained in the data. To this end, historical texts are digitized and processed by OCR or even transcribed manually. Due to the absence of standards, historical data often exhibits large variance, especially with regard to spelling. Hence, further processing either has to rely on fuzzy-matching strategies, or on standardization of the data.

In the Anselm project (Dipper and Schultz-Balluff, 2013), we opted for the second way. We provide normalized wordforms for the full corpus that have been manually annotated according to guidelines specifically created for this task (Krasselt et al., 2015). These normalizations can be useful for search queries, further downstream applications such as POS tagging, or as training data for automatic normalization methods.

This paper deals with means of quantitative evaluation of these normalization guidelines. We would like to quantify the degree of consistency that can be achieved with annotations according to the guidelines, i.e., the inter-annotator agreement (IAA). While a range of measures has been proposed for measuring agreement (e.g., see the survey by Artstein and Poesio (2008)), our task differs from common annotation tasks, such as part-of-speech tagging or semantic role labeling, in two important aspects: (i) the class of labels (the normalized wordforms) is open, and label distribution is sparse; and (ii) annotations are biased to be similar to the surface form of the token they belong to, and can match to different degrees. For example, we would like to score almost identical annotations like *nähme* – *nehme* 'take' (for the historical form *neme*) higher than annotations that are rather dissimilar, like *drückte* 'pressed' – *trocknete* 'dried' (for *trvckente*).

We investigate why conventional IAA measures are not suitable to the normalization task, and propose a new method that integrates common chance-corrected agreement measures, such as Fleiss's κ (Fleiss, 1971) or Krippendorff's α (Krippendorff, 1980). The novelty of our proposed method lies in the way the annotated wordforms are treated. First, we reframe normalization as a character-based task; and second, we model the inherent properties of normalization by mapping certain characters to more general categories.

We first present the annotation guidelines (Sec. 2) and the dataset that our evaluation is based on (Sec. 3). Sec. 4 discusses the problems that arise from applying common agreement measures to the normalization task. Sec. 5 introduces our new method, followed by an evaluation in Sec. 6, comparing and assessing the results of different ways of measuring agreement.

Proceedings of LAW X – The 10th Linguistic Annotation Workshop, pages 89–98,
Berlin, Germany, August 11, 2016. ©2016 Association for Computational Linguistics

2 Annotation of Language Changes

Languages evolve over time. This probably becomes most apparent in sound changes, which modify the way words are pronounced. In the long run, such changes are also reflected in the spelling of these words, cf. the pairs of word forms in (1), which are etymologically related, the ancestor being from Early New High German (ENHG, 1350–1650), the descendant from Modern German (MG).[1]

(1) a. *fr**iu**nd* / Fr**eu**nd 'friend' [N4]
b. ***ch**ind* / **K**ind 'child' [M1]

Of course, language evolution concerns all other linguistic levels as well, e.g. (2) shows changes in morpho-syntax (inflection).

(2) *vnser vrow**en*** (acc.sg.) / unser**e** Frau 'our lady' [M1]

Finally, words can change semantically or even get lost. In both cases, there is no direct, i.e. etymologically-related, equivalent in the modern language, see (3).

(3) a. *geitig* (geizig, lit. 'stingy') / gierig 'greedy' [M1]
b. *vnze* / bis 'until' [St2]

Since ENHG is already quite close to MG, it was decided to standardize ENHG forms to MG forms in the context of the Anselm project.[2] The question was now whether all the changes described above should be submitted to the same standardization procedure. For instance, if a word still exists in MG but with a different meaning (as in (3a)), should the word be replaced by the modern equivalent? What should be done with inflectional endings that have changed? After all, most inflectional differences would not hinder people from using and understanding the data, in contrast to clear semantic changes.

On the other hand, if we compare the effort it takes to automatically generate the forms, it is, of course, easier to generate forms that stay close to the original forms. However, for further use and processing of the data, forms are to be preferred in general that are maximally similar to modern data.

Ex	ENHG	Norm	Mod	Type
(1a)	*friund*	freund		
(1b)	*chind*	kind		
(2)	*vnser* *vrowen*	unser frauen	unsere frau	INFL INFL
(3a)	*geitig*	geizig	gierig	SEM
(3b)	*vnze*	unz	bis	EXT

Table 1: Normalization, modernization and modernization type of the examples (1)–(3) in the text.

ENHG	Norm	Mod	Type
da	da	als	SEM
er	er		
sein	sein	ihn	INFL
zum	zum		
dritten	dritten		
mal	mal		
verlaugnent	verleugnet	verleugnete	INFL
zuhant	zehant	sogleich	EXT
da	da		
kraet	kräht	krähte	INFL
der	der		
han	hahn		

Table 2: Normalization, modernization and modernization type of the sentence 'As he disowned him for the third time, the rooster crowed immediately' [Hk1].

2.1 Annotation guidelines

Rather than opting for one of the two forms, the guidelines designed in the Anselm project serve both camps by providing two levels of standardization, called *normalization* and *modernization*, see Krasselt et al. (2015). Normalization maps a given historical word form to a close modern (lower-cased) word form, considering sound and spelling changes. Modernization goes one step further and adjusts this form to an inflectionally or semantically appropriate modern equivalent, if necessary. In the annotation, modernized forms

[1] In the following examples, ENHG forms are given first, MG forms follow after the slash. The labels [N4], [M1] etc. refer to the text the example comes from, see Sec. 3.

[2] Another option has been traditionally pursued by researchers working on texts from the earlier period of Middle High German (MHG, 1050–1350). They standardized MHG word forms to an artificially-created, "idealized" MHG form, which is supposed to abstract from dialectal variation while keeping the "common" MHG characteristics.

Text	Tokens	Date	Dialect	Norm-Type			Mod-Type		
				ORIG	NORM	BOTH	INFL	SEM	EXT
Hk1	8,718	16th cent.	Central Bavarian	42.5	41.5	83.6	6.3	8.1	2.1
M1	10,274	14th cent.	Central Bavarian	41.3	40.8	82.1	8.4	7.4	2.1
N4	8,625	15th cent.	Alemannic + Bavarian	31.4	49.9	81.2	9.8	6.6	2.4
St2	8,873	14th cent.	Alemannic	32.9	53.1	86.0	4.4	6.8	2.8

Table 3: The texts of the four annotated fragments, with information about their provenance and frequencies (%) of normalization and modernization types.

are marked according to their type: INFL for inflectional modifications, SEM for semantically-determined replacements, and EXT for extinct ENHG word forms.[3]

Table 1 illustrates the two levels of standardization for the examples in (1)–(3), Table 2 shows the annotations for a short fragment of one text. If no morphological and/or semantic adjustment is necessary, the modernization and type levels are not filled.

3 Data

Our data comes from the Anselm corpus[4] (Dipper and Schultz-Balluff, 2013), a collection of texts from Early New High German (1350–1650). For the IAA evaluation, we selected fragments of 1000–1200 tokens of four manuscripts; see Table 3 for more information on these texts. All texts are written in dialects that are part of the language area called Upper German. Two of the texts are written in Central Bavarian but come from different centuries, 14th vs. 16th. The two other texts are from the neighboring region, Alemannic (with one of the texts also showing traits from Bavarian).

Table 3 also shows how many ENHG words are identical to MG words and do not need to be modified at all (column ORIG). The amount of "simple" normalizations, which only require sound and spelling adjustments, is shown in column NORM. The table also includes the frequencies of the different modernization types (columns INFL/SEM/EXT).

The four texts behave quite differently with regard to normalization and modernization. Judging from column ORIG, the two Alemannic texts, N4 and St2, seem more archaic than the two Bavarian ones, because they have a lower ratio of word forms that already correspond to MG. However, St2 has a very high ratio of words that can be normalized by adjusting the spelling only (column NORM). In fact, from a grammatical point of view, text St2 is the most modern one (see column BOTH). The fact that St2 shows the smallest proportion of INFL-type modernizations also points in this direction.

Of course, these figures do not tell us how difficult it is to normalize the individual texts. Common annotation errors are shown in (4) and (5); the examples first specify the original word form, followed by different normalizations as proposed by the annotators.

(4) Proper nouns

a. *iudas:* iudas, judas 'Judas'

b. *ysmahelite:* ismaeliter, ismaeliten, ismaheliten 'Ismaelis'

(5) Imperatives; subjunctive mood

a. *sag:* sag, sage 'tell'

b. *hoer/hoere:* hör, höre 'listen'

c. *neme:* nähme, nehme 'take'

There are also serious disagreements, resulting in semantically different words even on the normalization layer, as in (6) and (7). Very often, context information helps in disambiguating and, hence, avoiding such cases, so such disagreements are considerably less frequent than the cases above.

(6) Function words

a. *das:* das 'that' (pronoun), dass 'that' (conjunction)

b. *in:* in 'in' (preposition), ihn 'him' (pronoun)

[3] The guidelines define that extinct forms are standardized at the normalization level to forms that are compliant with reference lexicons, e.g. Lexer: `http://woerterbuchnetz.de/Lexer` or Deutsches Wörterbuch by Jacob and Wilhelm Grimm: `http://woerterbuchnetz.de/DWB`. In the Anselm corpus, Lexer was used as the reference lexicon.

[4] `https://www.linguistics.rub.de/comphist/projects/anselm/`

(7) Content words

a. *pin:* bin '(I) am', pein 'torment'

b. *dinen:* deinen 'your', dienen 'serve'

c. *holen:* hohlen 'hollow', höhle 'cave'

For the evaluation, passages in Latin and punctuation marks were removed from the texts, and all words were lower-cased. Five trained student annotators annotated these fragments. These annotations serve as the basis of the evaluation in Sec. 6.

4 Agreement Measures

The simplest way to measure agreement between annotators is "percentage agreement" ($agr_{\%}$), i.e., counting the number of items on which they agree and dividing the result by the total number of items. Percentage agreement has the drawback that it does not account for agreement *by chance.* A high chance agreement can occur, for example, when the annotation scheme only has a low number of distinct labels, or when certain labels occur much more often than others.

Therefore, most measures of agreement try to correct for chance. Two of the most widely-used agreement coefficients for nominal data are Scott's π (Scott, 1955) and Cohen's κ (Cohen, 1960), which both use the formula:

$$\pi, \kappa = \frac{A_o - A_e}{1 - A_e}$$

Here, A_o stands for observed agreement between two annotators, while A_e is the agreement expected by chance. Both coefficients estimate A_e from the distribution of the observed annotations in the evaluation data, the difference being that κ uses the *individual* distributions of each annotator, while π assumes an *identical* distribution for each.

Krippendorff's α (Krippendorff, 1980) is a similar, but more versatile coefficient. Like π, it assumes an identical distribution of labels, but is defined by the observed and expected *disagreement* between annotators:

$$\alpha = 1 - \frac{D_o}{D_e}$$

Despite this difference in definition, α and π are roughly equivalent (Artstein and Poesio, 2008, p. 567). The main advantage of α lies in the fact that it can use arbitrary *distance functions* to measure distance between labels. This allows for a more fine-grained treatment of disagreement than the binary "correct" or "wrong" distinction.

In the context of normalization, a possible distance function is *normalized Levenshtein distance (NLD),* which we define as follows:

$$NLD(a, b) = \frac{LD(a, b)}{\max(|a|, |b|)}$$

Here, $LD(a, b)$ is the Levenshtein distance between *a* and *b*, defined as the number of edits required to change *a* into *b* (Levenshtein, 1966), and $|x|$ is the character length of *x*. By using this function with Krippendorff's α, the disagreement between two annotations *a* and *b* effectively depends on their string similarity, with disagreements being considered less severe the more similar the two strings are.

It is possible to generalize π and κ to more than two annotators. Fleiss's κ (Fleiss, 1971) is a generalization of π, which we will call π^* here to avoid confusion. Krippendorff's α already accounts for multiple annotators.

4.1 Challenges for the Normalization Task

Normalization can be seen as a labelling task with nominal categories, where tokens are the annotation units, and normalized wordforms are the labels. This would allow us to use the aforementioned coefficients for calculating agreement. However, we believe that a naive application of these measures is not useful, and can even be misleading, for this task.

First, the set of all possible labels in the normalization task is the set of all morphologically well-formed words in the target language, of which only a small percentage will actually be seen in the annotated data. Estimating the label distribution from this data is therefore problematic, especially if the dataset is small. When calculating chance agreement, plausible alternative normalizations that do not occur in the training data will be given a probability of zero, which is not a realistic model.

Second, when the labels are words, most of the observed label types will usually be rare. Chance-corrected coefficients such as π/κ/α give more weight to rare labels than to common ones, which is usually desired (Artstein and Poesio, 2008). In the case of normalization, this seems unsound: we would expect the difficulty of agreeing on a normalization to depend mainly on the spelling characteristics and the closeness of the historical wordform to the modern target language, and not (or at least not exclusively) on its lexical frequency.

Third, using words as labels does not model the inherent property of normalization that most normalized wordforms will be similar, if not identical, to the historical token. When calculating chance agreement, all normalization candidates are considered equally, regardless of their similarity to the historical token. In other words, label probabilities are not conditioned on the items when calculating chance (dis)agreement for $\pi/\kappa/\alpha$. This is true for all annotation tasks, of course; however, for normalization, the large size of the label set exacerbates this problem.

A consequence of these factors is that a naive calculation of agreement will usually overestimate the annotators' performance. Particularly the second and third issue cause the expected chance agreement to be extremely low, while at the same time giving strong weight to almost any item where the annotators agree. The evaluation in Sec. 6 confirms these expectations.

5 Normalization as a Character-Based Annotation Task

Motivated by the problems discussed in Sec. 4.1, we explore the option of reframing the normalization task in the following way:

1. consider characters as the units for annotation instead of words; and
2. introduce an "identity" label for all normalizations where the character was not changed.

We will first describe how the mapping of annotations to characters is performed before discussing how this reframed task relates to the issues raised in Sec. 4.1.

5.1 Mapping Normalizations to Characters

Instead of considering words as our annotation units, we choose to view each character in the historical wordform as a unit of annotation. This raises the question of how to map word-level normalizations to individual characters, particularly if the historical and modernized wordforms are of different lengths.

Since normalizations derive from their original wordform by making adjustments to its spelling where necessary, and leaving other parts unchanged, this should be reflected in the character-based normalization by having identical characters line up if possible. We can achieve this by using the Needleman-Wunsch algorithm for sequence alignment (Needleman and Wunsch, 1970),[5] which favors aligning identical matches over any modifications or "gaps" in the sequences.

```
g e w a i n -    g e w a i n - -
g e w e i n t    - - w e i n t e
```

Figure 1: Character alignments using the Needleman-Wunsch algorithm

Units	Full		Diff	
	A	B	A	B
g	g	∅	_	∅
e	e	∅	_	∅
w	w	w	_	_
a	e	e	e	e
i	i	i	_	_
n	nt	nte	_t	_te

Table 4: Character-based representation of the token *gewain* being normalized as *geweint* (A) or *weinte* (B), showing either the full normalization (Full) or only the changes (Diff).

Figure 1 shows an example of the Needleman-Wunsch algorithm being used to align the historical wordform *gewain* to its potential normalizations *geweint* and *weinte* 'cried'. While this alignment has the desired property of lining up identical characters, we cannot use it directly because it introduces "gaps" in the historical wordform where characters are inserted—the annotation units should be fixed, though, regardless of the value of the normalization. We resolve this issue by merging insertions with the nearest non-insertion character to the left, with the (rare) exception of word-initial insertions, which are merged to the right. Table 4, column "Full" shows how our units and annotations look like after this process.

Finally, we introduce an identity label to represent matching characters. We do this before

[5] We use the Python implementation from the LingPy library (List and Forkel, 2016).

	Tokens	Word-based			Character-based		
		$agr_{\%}$	π^*	α_{NLD}	$agr_{\%}$	π^*	α_{NLD}
ALL	4558	0.9262	0.9254	0.9736	0.9698	0.9155	0.9184
MEDIUM	2858	0.8822	0.8804	0.9579	0.9551	0.9102	0.9138
STRICT	2673	0.9126	0.9112	0.9691	0.9653	0.9327	0.9355

Table 5: Inter-annotator agreement on normalization across five annotators; ALL = all tokens, MEDIUM = at least one annotator made a change to the original token, STRICT = all annotators made a change to the original token.

the merging step by replacing all identity alignments in the Needleman-Wunsch alignment with the identity label. The result can be seen in table 4, column "Diff". Note how this representation specifically highlights the *changes* made to the original token.

5.2 Advantages of the Character-Based Representation

Using character-based representations with identity labels does not completely solve the problems described in Sec. 4.1, but alleviates them significantly.

Instead of words, our label set now contains all possible character n-grams. While this is still a potentially unbounded set, the vast majority of labels are single characters only. This means that the effective size of our label set has been greatly reduced, allowing for a better estimation of the label distribution and reducing the "rare label" problem.

Introducing the identity label models the assumption that leaving characters unchanged is the "default" action. Under this assumption, the identity label will now be the most common label by far, and all other labels (representing modifications) will be comparatively rare. Since the agreement coefficients give more weight to rare labels, this means that agreement on actual modifications is now considered to be much more important than agreement on characters that do not change, which is exactly what we want.

Note that simply using the character-based representation *without* identity labels will overestimate the annotators' performance even more, since it greatly increases the number of units where the annotators agree. On the other hand, using identity labels directly on a word level does nothing to alleviate the issue of a potentially infinite label set.

6 Evaluation

We first compare agreement scores of the naive word-based evaluation with those obtained using the character-based representation of the task. For both scenarios, we calculate average percentage agreement ($agr_{\%}$) and Krippendorff's α using the NLD distance function defined in Sec. 4. We find that values for π and κ, either naively averaged over all annotator pairs or using the generalization of π^*, almost always differ only after the fifth or sixth decimal place; we therefore restrict ourselves to reporting π^*.

We evaluate separately on all tokens (ALL), tokens where at least one annotator made a modification to the historical token (MEDIUM), and tokens where *all* five annotators made a modification (STRICT).

Table 5 shows the agreement scores for this evaluation. The average word-based agreement over all tokens is 92.62%, and π^* values for the word-based task are always similar to the percentage agreement. Values for α_{NLD} are naturally higher, since it also considers partial agreement within the normalizations. For the character-based task, percentage agreement is always much higher, but π^* values are now noticeably lower compared to the percentage values. This is a consequence of the character-based reframing of the task being much more sensitive to agreement on the actual modifications (cf. Sec. 5.2).

Comparing the different evaluation sets, percentage agreement on the STRICT set is noticeably higher than on the MEDIUM set. This is particularly remarkable since the MEDIUM set only has 185 tokens more. Therefore, cases where annotators disagree whether a change to the historical wordform is even needed appear to be particularly problematic. On the other hand, if all annotators agree that a change needs to be made, they seem to reliably produce similar normalizations.

	Tokens	Word-based			Character-based		
		$agr_{\%}$	π^*	α_{NLD}	$agr_{\%}$	π^*	α_{NLD}
HK1	1157	0.9255	0.9247	0.9741	0.9701	*0.8957*	*0.9017*
M1	999	0.9252	0.9244	*0.9701*	0.9696	**0.9287**	**0.9322**
N4	1195	**0.9316**	**0.9306**	**0.9757**	**0.9712**	0.9239	0.9265
ST2	1207	*0.9221*	*0.9213*	0.9738	*0.9683*	0.9174	0.9186

Table 6: Inter-annotator agreement on normalization, separately for each text; highest score for each measure shown in **bold**, lowest score shown in *italics*.

This is supported even further by the fact that the STRICT set has the highest π^*/α_{NLD} scores in the character-based evaluation.

It is also interesting to compare the agreement by chance (A_e) between the two approaches. For π^*, the naive word-based evaluation has an expected agreement of $A_e^{\pi^*} = 0.0103$, which is not surprising considering that the pool of possible annotations is the set of all observed wordforms. For the character-based task, the majority of annotations are the identity label, which results in a high chance agreement of $A_e^{\pi^*} = 0.6312$. A better agreement between the annotators is therefore required to obtain a good π^* value.

For these reasons, we believe that the high agreement values of $\pi^* \geq 0.91$ on the character-based task provide stronger evidence for a good inter-annotator agreement on our dataset than the naive word-based evaluation does.

6.1 Per-Text Evaluation

Our evaluation dataset consists of passages from four different texts that exhibit different spelling characteristics (cf. Sec. 3). Since it is conceivable that this affects the difficulty of the normalization task, we also choose to evaluate on each text excerpt separately.

The results are shown in Table 6. Generally, there are only minor differences between the texts: for the word-based evaluation, N4 consistently shows the highest agreement, while ST2 usually has the lowest values (except for α_{NLD}, where M1 ranks worse). The same is true for $agr_{\%}$ on the character-based task. However, the agreement coefficients for the character-based task show very different trends: here, M1 gets the highest scores, while the values for HK1 are lowest by a noticeably margin.

This evaluation shows that our character-based evaluation is also useful for providing a different perspective on the annotated data than word-based agreement.

	Tokens	$agr_{\%}$	π^*
ALL	4558	0.8857	0.8171
MEDIUM	1230	0.5907	0.4681
STRICT	329	0.8839	0.8081

Table 7: Inter-annotator agreement on type of modernization; ALL = all tokens, MEDIUM = at least one annotator chose a modernization category (INFL/SEM/EXT), STRICT = all annotators chose a modernization category.

6.2 Type of Modernization

So far, the evaluation has focused on normalization alone. However, as described in Sec. 2, the annotation guidelines also include an additional modernization layer, which accounts for changes to the historical wordforms that go beyond spelling modifications.

Whenever annotators assign a modernization, they also need to select which type of adjustment they have performed. This allows us to evaluate agreement on the "type of modernization" they have chosen; we extend the three modernization types from our guidelines with two types for cases where no modernization has been performed, leaving us with these five categories: ORIG = no change from the original token; NORM = normalization, but no modernization; INFL = inflectional adjustment; SEM = semantic adjustment in the modernization; EXT = adjustment due to extinct wordform.

Table 7 shows that we achieve a reasonable agreement of $\pi^* = 0.8171$ on the assignment of these categories. However, restricting the eval-

	ORIG	NORM	INFL	SEM	EXT
ORIG	1452	11	20	36	1
NORM	–	2125	68	60	29
INFL	–	–	233	11	4
SEM	–	–	–	154	15
EXT	–	–	–	–	71

Table 8: Confusion matrix of annotator judgments between modernization types, averaged across all annotator pairs

	Tokens	$agr_{\%}$	π^*	α_{NLD}
ORIG	1357	1.0000	–	–
NORM	1930	0.9932	0.9870	0.9878
INFL	148	0.9715	0.9559	0.9606
SEM	63	0.8650	0.8453	0.8535
EXT	37	0.7694	0.7188	0.7227

Table 9: Inter-annotator agreement on modernization, using character-based evaluation, separately for tokens where all annotators agree on the type of modernization.

uation to tokens where at least one annotator chose one of the actual modernization categories (INFL/SEM/EXT; row MEDIUM in Table 7) results in a very low score of 0.4681. A further restriction to tokens where *all* annotators chose one of these categories results in a much better score again, however, this was only the case for 329 tokens. These results show that our annotators disagree strongly on when to actually assign a modernized wordform at all; in the few cases where they all agree that a modernization has to be assigned, the agreement on the type of modernization is reasonably good.

To further illustrate this point, Table 8 shows a confusion matrix on modernization types. For each of INFL/SEM/EXT, the second most often selected category by another annotator was NORM, i.e., a normalization where no additional modernization was performed. However, disagreement within these categories of INFL/SEM/EXT occurs only rarely, confirming the interpretation of the values in Table 7. Also, confusion with the ORIG category is also comparatively rare, showing that wordforms which do not need to be changed are much less problematic.

6.3 Character-Based Evaluation of Modernization

Due to the nature of the modernization layer, a character-based evaluation of the wordforms is problematic, since modernized forms usually do not need to bear any resemblance to the historical token. An exception are modernized forms that have been assigned due to inflectional changes (INFL), which we would assume to be similar to the respective historical and normalized forms.

To test this assumption, we evaluate character-based agreement on the modernization layer for tokens where all annotators agree on a modernization type (Table 9). For ORIG and NORM, we assume the modernized wordform to be identical to the normalization. The results confirm our expectations: π^* on INFL is 0.9559, while it drops considerably for SEM and EXT; however, the significance of these results might be limited due to the low sample size for these cases.

Another notable result is the extremely high agreement ($\pi^* = 0.9870$) for tokens where all annotators agree on type NORM. This tells us that most of the disagreements from the normalization evaluation (cf. Table 5) stem from cases where at least one annotator decided that a modernization was necessary; these tokens therefore appear to be more difficult to agree on not only on the modernization layer, but already on the normalization layer.

While it is plausible that extinct wordforms, as well as words with different meaning or inflection than in modern language, are inherently more difficult to annotate, the intention of the guidelines was to move this difficulty to the modernization layer, while having unambiguous rules for the annotation of the normalization layer. These results show that while we achieve a good reliability overall, the guidelines were not able to remove this difficulty completely for these cases.

7 Discussion

In this paper, we presented and evaluated a method to measure inter-annotator agreement on normalization of historical data. We argue that our character-based evaluation approach is more appropriate for this task from a theoretical perspective, and showed that it behaves differently than a naive word-based measure.

We have found that the scores resulting from our method correspond well to our intuitive judgments. As a direction for future research, it would be useful to conduct a systematic evaluation of this notion. For that purpose, human annotators would rate normalizations for agreement, and the level of correspondence would be revealed by how well the metrics can reproduce the rankings of the human annotators. However, the rating of normalizations is not in itself a trivial task. It would also have to be based on entire texts rather than isolated pairs of normalizations, since expected agreement cannot be calculated for isolated pairs and, hence, a comparison with our scores would not easily be possible. For these reasons, we did not conduct such a study for this paper.

Our proposed method is certainly not the only way to accomodate the specific properties of the normalization task. Instead of viewing the task on a character level, normalizations could also be seen as sets of edit operations on a word. This can easily be derived from the Needleman-Wunsch alignment that we already use (cf. Fig. 1): instead of the normalization *geweint*, we could define the annotation of the token *gewain* to be a set of edit operations $\{4\colon a \rightarrow e, 6\colon n \rightarrow nt\}$, and use a set-based agreement measure on it—see, e.g., Passonneau (2004) for a set-based measure applied to coreference annotation. However, this approach is also not free of problems: in the annotated set, the position of edit operations is important, but for purposes of calculating chance agreement, positional information should not be included. While we believe this difficulty can probably be resolved, we did not explore this option further.

We are aware of only one approach that reports agreement figures on the task of normalizing historical data, Scheible et al. (2011), who deal with data from Early Modern German (1650–1800) and report word-based percentage agreement of 96.9%. As we have argued, word-based evaluation alone cannot adequately assess performance of the annotators because partial agreement is not considered, and also this measure does not try to correct for chance.

Normalization is also sometimes performed on other types of data, such as dialectal or social media texts. Our method of evaluating IAA can be generalized to these datasets as long as it is sensible to frame them as a character-based annotation task, i.e., the annotation values should be derived from (and typically be similar to) the surface forms of their respective tokens. The same considerations apply when transferring this approach to other open-class annotations, e.g. lemmatization.

Acknowledgments

We would like to thank Julia Krasselt and all of our student annotators for their help in creating the evaluation dataset. The research reported here was supported by Deutsche Forschungsgemeinschaft (DFG), Grant DI 1558/4.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Stefanie Dipper and Simone Schultz-Balluff. 2013. The Anselm corpus: Methods and perspectives of a parallel aligned corpus. In *Proceedings of the NODALIDA Workshop on Computational Historical Linguistics*, Oslo, Norway.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Julia Krasselt, Marcel Bollmann, Stefanie Dipper, and Florian Petran. 2015. Guidelines for normalizing historical German texts. *Bochumer Linguistische Arbeitsberichte*, 15.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. SAGE, Beverly Hills, CA.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Johann-Mattis List and Robert Forkel. 2016. LingPy. A Python library for historical linguistics. Version 2.4. `http://lingpy.org`. With collaborations by Steven Moran, Peter Bouda, Johannes Dellert, Taraka Rama, Frank Nagel, and Simon Greenhill: Max Planck Institute for the Science of Human History.

Saul B. Needleman and Christian D. Wunsch. 1970. A gene method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453.

Rebecca J. Passonneau. 2004. Computing reliability for coreference annotation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, volume 4, pages 1503–1506, Lisbon, Portugal.

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011. A gold standard corpus of Early Modern German. In *Proceedings of the ACL-HLT 2011 Linguistic Annotation Workshop (LAW V)*, pages 124–128, Portland, Oregon, USA.

William A. Scott. 1955. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19(3):321–325.