Representation Problems in Linguistic Annotations: Ambiguity, Variation, Uncertainty, Error and Bias

Christin Beck[1], **Hannah Booth**[1,3], **Mennatallah El-Assady**[2], and **Miriam Butt**[1]
[1]Department of Linguistics
[2]Department of Computer Science
University of Konstanz, Germany
[3]Department of Linguistics
Ghent University, Belgium
firstname.lastname@uni-konstanz.de

Abstract

The development of linguistic corpora is fraught with various problems of annotation and representation. These constitute a very real challenge for the development and use of annotated corpora, but as yet not much literature exists on how to address the underlying problems. In this paper, we identify and discuss five sources of representation problems, which are independent though interrelated: ambiguity, variation, uncertainty, error and bias. We outline and characterize these sources, discussing how their improper treatment can have stark consequences for research outcomes. Finally, we discuss how an adequate treatment can inform corpus-related linguistic research, both computational and theoretical, improving the reliability of research results and NLP models, as well as informing the more general reproducibility issue.

1 Introduction

Linguistically annotated corpora have for many decades occupied a firm place in the linguistics toolbox. They are of vital importance for theoretical and computational linguistic research in providing empirical evidence for language use, both from a qualitative and quantitative perspective. Moreover, machine learning algorithms typically applied in the context of Natural Language Processing (NLP) require annotated datasets for building language models. Annotated corpora now exist in many forms and much effort has been devoted to developing specific schemes for different levels of linguistic analysis (e.g., phonology, morphosyntax, semantics, pragmatics). For example, the Penn Treebank (Marcus et al., 1999) and the Universal Dependency Treebank (e.g., de Marneffe et al., 2014) are syntactically annotated corpora, while, e.g., discourse relations are annotated in the Penn Discourse Treebank (Webber et al., 2019). Although many levels of linguistic annotation can be performed via automated means, some manually annotated training data is typically required as a foundation. Moreover, there are certain types of corpora where manual annotation remains the best option in the face of complex linguistic phenomena (e.g., historical language stages or non-standard varieties).

Any manual annotation process represents a compromise between an accurate linguistic analysis and an annotation scheme which is generalizable enough to serve computational tools and the end user. Annotation schemes are typically designed with a specific purpose in mind, and this will bear heavily on the decisions made. The compromises tend to revolve around details of the annotation scheme and one possible result of hard-fought compromises is that the resulting representations may actually be inaccurate with respect to several factors. In this paper, we identify five major factors: (i) ambiguity, (ii) variation, (iii) uncertainty, (iv) error, and (v) bias (see also Figure 1). In particular, uncertainty has already been recognized as a problem in linguistic corpora (Jurgens, 2013; Cassidy et al., 2014), with a special focus on the interrelation between linguistic ambiguities and uncertainty in the annotation of historical linguistic data (Seemann et al., 2017; Merten and Seemann, 2018).

In this paper we extend the discussion beyond ambiguity and uncertainty to include variation, error and bias as sources of representation problems in linguistic annotations and argue for the importance of

The 14th Linguistic Annotation Workshop, pages 60–73
Barcelona, Spain (Online), December 12, 2020.

developing a robust framework which explicitly treats these problems. We focus primarily on representation problems in historical corpora, but the set of problems is transferable to other types of linguistically annotated resources (Chambers et al., 2014; Plank et al., 2014; Pavlick and Kwiatkowski, 2019). As part of future work, we intend to build a computational implementation that is based on the crucial foundations laid out in this paper.

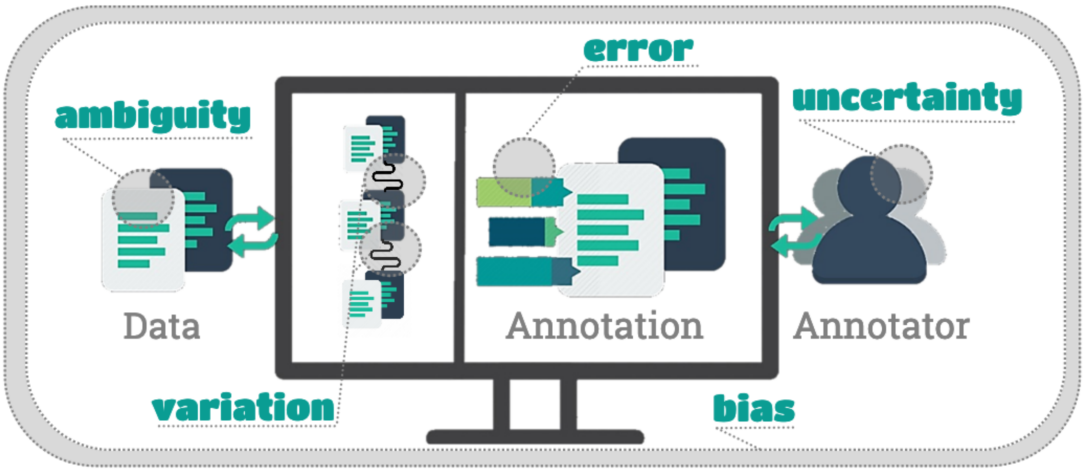

Figure 1: Representation problems in linguistic annotations come from five distinct sources: (i) **Ambiguities** are an inherent property of the *data*. (ii) **Variation** is also part of the *data* and can, e.g., occur across documents. (iii) **Uncertainty** is introduced by an *annotator*'s lack of knowledge or information. (iv) **Errors** can be found in the *annotations*. (v) **Biases** are a property of the *complete annotation system*.

2 Background and Related Work

Existing approaches typically treat representation problems in one of three ways in linguistic annotation processes: (i) stochastic treatment, (ii) assignment of an 'other/miscellaneous' category, (iii) left unannotated. In Part-of-Speech (POS) tagging and syntactic parsing, ambiguities are often treated in a stochastic manner, so that among the possibilities the option which is most likely is chosen (cf. Most Frequent Class Baseline; Jurafsky and Martin, 2009). In this way, each token receives a single tag and the ambiguity property is lost altogether. For example, Dipper et al. (2013) develop a POS tagging policy for historical German where ambiguous material receives the tag of the more frequent or of the historically older usage. Another approach is to mark entities about whose interpretation an annotator is uncertain with a specific tag ('other/miscellaneous' category), signaling that no adequate annotation is available. For example, in the Corpus of Historical Low German (Booth et al., 2020), clauses which are ambiguous between matrix (`IP-MAT`) and subordinate status (`IP-SUB`) are tagged as `IP-X`. A non-historical example comes from TimeBank-Dense (Chambers et al., 2014; Cassidy et al., 2014) where ambiguous temporal relations are tagged as 'vague'. Although this captures some level of ambiguity and uncertainty, it does not necessarily allow for an adequate representation of the underlying problem, information which is potentially of high interest to the end-user. A third option often employed is to leave uncertain material unannotated. In this case, the respective pieces of data are typically set aside and do not play a role in linguistic investigations or NLP downstream tasks.

Some efforts have been made towards more sophisticated schemes for explicitly marking representation problems. For instance, Merten and Seemann (2018) have developed a novel interface which enables an annotator to capture different sources of uncertainty while annotating POS and specific syntactic constructions in a historical corpus of Middle Low German (see also Seeman et al., 2017). Their scheme allows annotators to capture uncertainty via a tag indicating one of three types: (i) category A is more likely than B, (ii) A and B are equally likely, (iii) unsure. Lüdeling (2017) has proposed a corpus annotation model which captures variation, e.g., in word pronunciation, by explicitly labeling the type of variation and the variant involved in an additional annotation layer. Similarly, Dipper et al. (2013) capture

spelling variation in their historical tagset by providing information on three levels: (i) the original token (diplomatic level), (ii) a modernized version of the token (tokenization level), and (iii) a tag for the type of variation (tag level). Likewise, Barteld et al. (2014) have proposed a multi-level annotation approach for incomplete language phenomena, i.e., ambiguity, underspecification or uncertainty.

In other contexts, uncertainty has been measured via various types of scale. For example, Jurgens (2013) employs a Likert-scale based approach for annotating ambiguous word senses, where the possible senses are rated on a numerical scale with respect to the likelihood of their occurrence in specific contexts. Vashishtha et al. (2019) use confidence ratings ranging from 'not at all confident' to 'absolutely confident' when annotating temporal relations between pairs of events. Zhang et al. (2017) let annotators specify whether an inference relation between sentences is 'very likely', 'likely', 'plausible', 'technically possible' or 'impossible' while annotating natural language inferences (NLI). Also in the context of NLI, Pavlick and Kwiatkowski (2019) make use of a sliding bar, which contains numerical values which range from indicating that an inference relation is 'definitely true' to indicating that it is 'definitely not true' to capture human judgements. Additionally, Pavlick and Kwiatkowski (2019) assess annotator disagreement as a measure of uncertainty by modeling these judgements as distributions instead of aggregated scores. Similarly, Chen et al. (2020) let annotators give judgements about the likelihood of inference relations via a sliding bar, and model 'uncertain natural language inferences' via scalar regression, predicting the probability at which a premise entails a hypothesis. Passonneau and Carpenter (2014), on the other hand, present a probabilistic model based on maximum likelihood estimation for annotating word senses, which gives a confidence estimate for each annotation label as a measure of certainty. Plank et al. (2014) capture uncertainty by measuring annotator disagreement on the basis of inter-annotator $F1$-scores and the confusion probability between annotators. In addition, Plank et al. (2014) show that by incorporating information about annotator disagreement into the training of an NLP-model for POS-tagging, relevant downstream tasks, i.e., Named Entity Recognition and Chunking, can be improved.

Although these approaches take up the problem of annotation and representation problems, they do not work out a generally applicable framework. Additionally, it is still unclear how the factors we identify as major problem sources interact throughout the various phases involved in corpus development and use. We see the important barrier to overcome here as a conceptual one, in the sense that concepts like 'uncertainty' and 'ambiguity' are often used interchangeably, despite there being inherent differences between the sources of these representation problems. Understanding these sources on their own terms, as this paper proposes, is a crucial prerequisite for developing more adequate treatments and in turn more reliably annotated corpora.

3 Phases of Corpus Development and Use

Developing an annotated corpus for linguistic research involves a number of work steps which are crucial for successful corpus design. The corpus development workflow consists of data selection and processing, typically including digitization, normalization and automatic pre-processing, and cycles of annotation. We see two major parts of a typical corpus development process: Data Selection and Processing (Phase I) and Annotation (Phase II). Following the corpus development, there is a third part, Interpretation (Phase III), which pertains to corpus use, addressing the issue of interpreting the annotated data.

3.1 Phase I: Data Selection and Processing

Prototypically, corpora consist of several machine-readable text files. Finding appropriate texts for a corpus is not a trivial task and data selection, i.e., text collection, is the first fundamental step in corpus construction. Ensuring a balanced corpus is a high desideratum, though often not achievable, particularly with respect to historical corpora, since often only a limited number of texts from particular domains are available. Data processing usually comprises digitization of any non-digitized textual material, spelling normalization and an automatic pre-processing (e.g., POS tagging, shallow syntactic parsing), and prepares the data for manual annotation by enriching the raw text with basic linguistic information.

3.2 Phase II: Annotation

Some corpus building stops after Phase I, because the amount of annotation is already sufficient for the intended task. Other corpora are enriched with further, often manual, annotations capturing more complex linguistic phenomena. The manual annotation process generally consists of iterative cycles of annotation, evaluation and error correction. At times, the manual annotation process is sped up by combining it with machine learning via successive cycles of manual annotation, training, automatic annotation and corrections. The annotation task itself can be broken down into two subtasks: identification and classification. First, the linguistic unit which is to be annotated has to be identified. The classification task then deals with assigning an annotation label to the previously identified linguistic unit. Manual annotation pipelines will also be guided to some extent by the specific tool which is used to conduct the annotation, of which a range exist, e.g., Annotald (Beck et al., 2015) and WebAnno (Eckart de Castilho et al., 2016).

3.3 Phase III: Interpretation

Annotations are usually informed by extensive theoretical analysis of linguistic structures. In turn, annotated corpora are essential for linguistic research, since they provide empirical evidence for a variety of linguistic phenomena, with end-users generating relevant insights for linguistic theory. In this way, the interpretation of the annotated values is crucial to future scientific developments. Moreover, corpora are a necessary prerequisite for the building of language models within NLP. Such models are based on interpretations (learning) of the annotated data. Thus, adequate representations of the linguistic data are vital for end-users of corpora. Without explicit treatment, representation problems often persist in post-hoc analyses of the annotations, rendering linguistic findings and computational models potentially unreliable or even misleading. In the next section, we characterise the various sources of representation problems, how they surface at the different phases, and show how they merit a more explicit treatment.

4 Sources of Representation Problems

We identify five sources of representation problems in linguistic resources, which we explain in this section: (i) ambiguity, (ii) variation, (iii) uncertainty, (iv) error and (v) bias. They are relevant at all three phases of corpus development and use, and interact with one another in a complex fashion.

4.1 Ambiguity

Ambiguity is an inherent property of natural language and therefore also of corpus data, see Figure 1. Ambiguity occurs whenever an entity in principle allows for multiple interpretations. Ambiguities between form and meaning appear frequently in natural language at all linguistic dimensions (e.g., phonological, morphosyntactic, lexical or pragmatic) and are a key source of representation problems. For our purposes, we define ambiguity as any instance where linguistic material in principle allows for more than one interpretation.

We propose to capture the extent to which an instance of linguistic ambiguity can be resolved (e.g., via contextual cues and/or world knowledge) via three broad categories: (i) ambiguity can be fully resolved; one interpretation, (ii) ambiguity cannot be entirely resolved, but a preference can be expressed for one interpretation; multiple interpretations, relatively ranked, and (iii) ambiguity cannot be resolved, and no preference can be expressed for any interpretation; multiple interpretations, equal ranking (see also Merten and Seemann (2018)). Ambiguity poses particular problems for representation at Phase II (annotation), leading to challenges in the identification and classification of linguistic units.

Many classic examples constitute *class ambiguity*, referring to cases where a linguistic unit allows for more than one classification. A well-known example from English concerns gerunds in *-ing*, which are often assumed to be a 'mixed category' between noun and verb (Hudson, 2003; Malouf, 1996).[1] However, not all *-ing* forms are equally unresolvable, see example (1), taken from Lowe (2016, 402). In (1-a), *missing* exhibits properties exclusively associated with nominals (adjectival premodification,

[1]This type of class ambiguity poses a problem to POS tagging, which we broadly described as a Phase I process. But POS tagging is at times part of the manual annotation (Phase II) and more generally feeds into the parsing of larger syntactic units.

PP complement), while in (1-b) the properties of *missing* are entirely verbal (adverbial modification, bare logical subject, bare object). In other words, a potentially ambiguous form can be resolved to one of the possible interpretations. The unresolvable ambiguity arises in contexts like (1-c), where *missing* exhibits both nominal properties (logical subject is a possessive phrase) and verbal properties (adverbial modification, bare object). This would be a case of two interpretations (N and V) with an equal ranking. Furthermore, the gerund example highlights a more general issue which feeds class ambiguity, i.e., that many of the categories widely recognized in linguistics and thus implemented in annotations are typically encoded through bundles of properties. As such, it is expected that there will be clear-cut cases, as (1-a)-(1-b), but also items whose properties indicate membership of more than one category, as (1-c).

(1) a. [His stupid **missing** of the penalty] lost us the game. (unambiguously nominal)
b. [Him stupidly **missing** the penalty] lost us the game. (unambiguously verbal)
c. [His stupidly **missing** the penalty] lost us the game. (ambiguously nominal/verbal)

An example of ambiguity which is not fully resolvable, but where a preference *can* be expressed for one of the possible interpretations is shown in (2), where *crane* is ambiguous between the bird-type (animate) and the machine-type (inanimate). In (2), the preceding context does not point strongly towards a preference, but the subsequent context indicates the machine-type interpretation to be the most likely, based on the world knowledge that machine-type cranes are involved in apartment building.

(2) On the river side, I saw a bunch of **cranes**. The new apartments are starting to look really nice.

For this example, the ambiguity does not concern POS tags but would be relevant in a resource where one is annotating for animacy, e.g., VerbNet (Kipper Schuler, 2005).

Moreover, class ambiguity can result from processes of historical language change, e.g., grammaticalization (Hopper and Traugott, 2003). A classic example is the development whereby a demonstrative becomes a complementizer via grammaticalization (Heine and Kuteva, 2002), as exhibited with English *that*, e.g. (3), where the demonstrative function remains ('persistence'; Hopper, 1991).

(3) a. I say **that**: there is a problem. (demonstrative with cataphoric reference)
b. I say **that** there is a problem. (complementizer)

Without punctuation and prosodic cues, the form *that* exhibits class ambiguity between demonstrative and complementizer. This is particularly relevant for historical language stages where one typically only has access to written texts, and thus to no prosodic information.

Boundary ambiguity refers to instances where a surface string can in principle be segmented into smaller units in more than one way, resulting in alternative boundary divisions. This is related to the identification part of the annotation task. An example from syntactic constituency is provided in (4), where *the man with the telescope* can be one larger nominal constituent with internal modification (embedded PP), see (4-a), or two separate constituents (NP PP), see (4-b).

(4) Mary saw **the man with the telescope**
a. Mary saw [the man [with the telescope]]
b. Mary saw [the man] [with the telescope]

Moreover, such examples present a further type of ambiguity, *attachment ambiguity*, which is relevant for identifying relations between segments and arises when there is more than one possible interpretation of these relations. In (4), the PP *with the telescope* can in principle attach at more than one level in the phrase-structure: in (4-a), it attaches at NP-level; in (4-b), it attaches at VP-level. Just as grammaticalization can feed class ambiguity, reanalysis as a mechanism of change (de Smet, 2009) is related to boundary and attachment ambiguity, in the sense that a surface string is assigned a new bracketing interpretation, typically via ambiguous 'bridging contexts' (Heine, 2002).

Since ambiguity is inherent to the data, it can also occur as a representation problem at Phase I. For instance, word and sentence segmentation can be the locus of ambiguities, particularly in historical

corpora. In many modern languages, word and sentence boundaries can be identified on the basis of white spaces and punctuation. In older handwritten manuscripts and early printed sources, however, the use of spaces and punctuation can differ quite substantially from the modern usage in terms of functionality (Dipper et al., 2013). For example, while in modern German words represent syntactically meaningful units, Old High German scribes often employed separation in the form of spaces or the absence thereof to group words into prosodic units (Fleischer, 2009). Thus white spaces can be ambiguous with respect to their linguistic function.

Although ambiguity is an omnipresent problem in the corpus development process, it is in most instances not captured by the annotations. Instead, the general practice is to stochastically determine and use the most probable annotation label, losing the ambiguity property altogether. If ambiguity does not receive adequate treatment, then this has consequences for the user at Phase III. If the ambiguity is not captured at all, then interpretations may be simply false, and indeed rich information will be lost which is often very relevant to linguistic investigations and likewise to computational language models.

4.2 Variation

Variation is when a particular variable is expressed via multiple variants. The variants are in principle interchangeable so that there are no (structural) conditions excluding one of the variants (Lüdeling, 2017). Variation may be conditioned by factors which are extra-linguistic, e.g., time period, dialect, genre, author/speaker of text, or linguistic factors such as language change or the linguistic environment a variable occurs in.

The dative alternation in English is an example of variation (Bresnan et al., 2007; Lüdeling, 2017). The dative alternation refers to the availability of two different (syntactic) dative constructions in English, which can be used interchangeably while expressing essentially the same meaning. The variants are illustrated in (5), with (5-a) showing the variant which contains a prepositional dative structure (NP PP) and (5-b) showing the double object variant (NP NP).

(5) a. Mary gave [an apple] [to John].
b. Mary gave [John] [an apple].

The variation here is determined by a conglomeration of different linguistic factors, including animacy, discourse accessibility (givenness/information structure) and weight (Bresnan et al., 2007).

Like ambiguity, variation is an inherent part of natural language and thus of the corpus data (see Figure 1). Therefore, variation constitutes a representation problem at all three phases of corpus development and use. In contrast to ambiguity, it is not an annotation issue in terms of identification and classification, but rather that a single interpretation (variable) manifests itself in two or more (variants). Sometimes, the variants are captured in an annotation scheme, but not necessarily linked to a single variable (e.g., Dipper et al., 2013; Lüdeling, 2017). Although such a treatment provides significant information about the variant, the variation as a whole cannot easily be harnessed without prior knowledge of the precise character of the variation, since crucial information about the other variant(s) is not easily accessible.

In other treatments, the variants are levelled out and expressed as a single interpretation (either as one of the variants or as a generalizing variable), in which case the variation property is lost overall. For example, normalization is a process at Phase I which leads to a levelling out of spelling variation, where one variant is favored over another. Spelling variation occurs across the board in historical texts, reflecting e.g., dialectal and/or temporal differences, as orthographic norms are a trait of modern times. For example, Bollmann et al. (2014) find the three dialect variants *chind, kínt, kynt* for 'child' in Early New High German, while *kind* is used in (late) New High German. In such cases, normalization is often a necessary means to an end. Normalization enables researchers to leverage existing algorithms and tools for text processing, facilitating, e.g., POS tagging (Bollmann et al., 2014). Furthermore, regularized spelling has the advantage of facilitating keyword and n-gram searches in corpora (Kytö, 2010). Yet, normalization can also lead to a crucial loss of information, since spelling variation may provide valuable linguistic insights relevant for annotation and interpretation.

A linguistically relevant instance of variation which could be levelled out by normalization is the

variation between multi- and single-word spellings caused by univerbation (Dipper et al., 2013; Lüdeling, 2017). Univerbation is an instance of language change whereby multiple words which form a fixed expression are reanalyzed as one word, with an intermediate stage where both variants are used interchangeably. For instance, in the Penn Parsed Corpora of Historical English (Santorini, 2010) both *nevertheless* (single-word) and *never the less* (multi-word) occur. Keeping the spelling variation intact provides insights into where a particular text is situated on the trajectory of a change (Dipper et al., 2013) and potentially reveals the linguistic factors which led to the development of the fixed multi-word expression. Still, without a more explicit treatment of the variation, together with the relevant a priori knowledge, this is hard to explore.

Another reason why variation merits particularly nuanced treatment in historical corpora is because certain historical processes feed the issue, e.g., grammaticalization and competition. As mentioned, grammaticalization involves a change whereby a particular form takes on a new function. This in turn can result in variation, if a marker of this particular new function already exists in the language. Secondly, competition between variants of a single variable often results in decreased variation in a particular corner of the linguistic system, whereby one variant outcompetes another (Kroch, 1989; Pintzuk, 2003).

Overall, losing variation significantly impairs a language resource in terms of its accurate reflection of the linguistic characteristics of the data. As with ambiguity, this loss of information is not insignificant, since variation is another linguistically relevant property of language which is often of prime interest to the user (Labov, 1994; Tagliamonte, 2006; Chambers and Schilling, 2013).

4.3 Uncertainty

Uncertainty arises wherever multiple possible interpretations of data present themselves, but the relevant knowledge or information to unequivocally opt for one of the interpretations is not available (Bonneau et al., 2014). Uncertainties can be part of the process of data selection and processing. For instance, corpus developers might be uncertain about which texts fit best with the objective of the corpus and which parameters to choose for text processing. In addition, the NLP tools employed for pre-processing can introduce uncertainty (John et al., 2017). This is an issue particularly for historical corpora, since, e.g., POS taggers are often trained on data from more recent time periods, given that the necessary amount of annotated training data for the historical period is typically unavailable. This renders the tagging results on historical data potentially unreliable, which in turn leads to uncertainty.

Moreover, uncertainties occur frequently in the annotation phase, as depicted in Figure 1. An issue which arises with historical data is that the crucial knowledge for the annotation of a specific historical language structure may not yet have been generated. A further problem is that human annotators cannot function as native speakers of a historical language. Due to incomplete knowledge, annotators may not be able to readily identify and interpret a given structure, and may therefore be uncertain.

A further uncertainty is caused by the annotations themselves. Linguistic annotations instantiate theory to some degree, focusing on some phenomena over others, with many phenomena not yet studied in much depth. The corresponding code books or manuals for annotation therefore hardly ever tend to be comprehensive and complete before the begin of the annotation process and are necessarily extended and changed as part of a cyclical annotation process (Hovy and Lavid, 2010). This results in uncertainty as to how unanticipated phenomena should be annotated, new tags defined, and how phenomena not covered by established research should be treated (Hovy and Lavid, 2010). The uncertainties that are encountered at Phase I and II are marked as such and made transparently explicit to the end user only rarely, thus persisting into Phase III (interpretation).

4.4 Error

In addition to the sources already discussed (ambiguity, variation and uncertainty), errors may also occur in linguistic annotations as representation problems (see Figure 1). At Phase I, errors can already be present in the data sources themselves. For example, with respect to historical manuscripts, scribal errors are common, particularly in texts which have been copied multiple times by different scribes (Penzl, 1967; Neidorf, 2013). Moreover, the source texts are often not in good repair, with stains on the paper and damaged pages potentially producing digitization errors. Texts can be digitized by either hand-keying or

scanning via OCR (optical character recognition) software. Handkeying has the advantage of generally being more accurate (though not error-free), but is time-consuming and might not be suitable when dealing with a large number of texts. OCR systems, on the other hand, generally work fast and are able to handle large quantities of data. Yet, the scripts, characters and diacritics of historical manuscripts and early printed texts are often challenging for OCR systems, requiring a non-trivial amount of post-processing, including error identification and correction (Boschetti et al., 2009; De Simone et al., 2018; Schulz and Kuhn, 2017). The process of identifying and correcting the erroneous text passages is laborious, produces a high cognitive workload and requires expert philological knowledge. The resulting corpus might therefore contain errors produced by the OCR system, but these will not necessarily be distinguishable from errors caused by human unsystematicity. These errors might have an impact on subsequent processes, e.g., POS tagging. Moreover, the automatic pre-processing steps are prone to errors themselves, which are often not transparent to annotators and users.

Manual annotation at Phase II is time-consuming and cognitively heavy. Human errors might therefore occur and not be detected, even in iterative rounds of annotation and correction. This is especially relevant for historical corpora, since human annotators lack native-speaker competence, as well as the cultural and pragmatic knowledge of the historical language stage, and may at times be unable to analyze certain linguistic structures accurately. Moreover, it is not always possible for the data to be annotated by several annotators, since the annotation of complex linguistic phenomena often requires expert knowledge. Again, this particularly applies to historical corpora, where only a few trained researchers with the relevant knowledge may exist. In this way, calculating inter-annotator agreement (e.g., via Cohen's kappa (Cohen, 1960), Krippendorf's alpha (Krippendorff, 2004) or inter-annotator $F1$-scores (Plank et al., 2014)) is not possible and significant errors may remain undetected.

Since revision and correction of annotations are costly procedures, it is often the case that a version of the corpus is already published after the first round of annotation, to provide the research community with the data as soon as possible. Errors are then usually reported by the community and the corpus developers can in principle react to this by re-annotating the data for the next release. However, in practice different sites often end up maintaining different versions of the corpus or researchers might 'curate' the data themselves, working with their own versions. This impedes reproducibility and may lead to imperfect research results.

4.5 Bias

Bias represents an influence which leads to a preference or tendency for one thing over another. Often, neither the end-user nor the annotator may be conscious of such biases, which can produce representation problems in every phase of corpus development and use.

In general, corpora should be representative and balanced (Leech, 1991; Gries and Berez, 2017). A corpus can be representative of a specific genre, register or variety, representing the targeted subgroup via the text samples contained in the corpus. A corpus is balanced (unbiased) when the size of the subsamples, i.e., the samples of different genres, registers, or varieties, is proportional to the size of the subgroups which the corpus aims to represent. With respect to diachronic corpora, genre imbalance is rather the norm than the exception, which in turn leads to a sampling or selection bias. While a large amount of textual data is usually available for more recent time stages of a language, the data for the longer standing past is generally scarce. The historically older and sparser data is generally less diverse, consisting of fewer genres, registers and text types (Gippert and Gehrke, 2015). It is often the case that all available historical texts for the relevant time periods are included in a corpus to be able to cover the diachrony of a language as much as possible (Reppen, 2010). Genre imbalances across time periods can hinder comparability over time, which is the core remit of diachronic investigations. Furthermore, text processing can be subject to a bias, since the parameters and tools chosen influence the shape of the resulting data. Søgaard et al. (2014) point out that language technology is generally biased towards English newswire (selection bias), with better overall performances of NLP tools on English newswire data than on any other text genre and other lower-resourced languages.

At Phase II, several more biases can occur. For one, the annotator might already have a theory in mind

about the phenomenon to be annotated, which has not been empirically evidenced, imposing a theory bias on the data. Moreover, comparative fallacy (Merten and Seemann, 2018), e.g., misinterpretations arising through comparing the historical language with one's own native language, might lead to substantial biases. For another, a learning effect might occur during the course of annotation, biasing the resulting data, in the sense that the resource will not necessarily be internally consistent.

4.6 Interrelations between Sources

One type of representation problem rarely occurs on its own, with one type of problem often leading to another. *Ambiguity* interacts strongly with uncertainty, but it is crucial to differentiate between the two: while ambiguity is inherent to the data, a human annotator/user or an algorithm might be uncertain when multiple interpretations of the data present themselves, as sketched in Figure 1. Similarly, *variation* might produce uncertainty if the precise nature of the variants involved and/or their conditioning factors cannot be easily recovered (lack of knowledge/information). Moreover, the human might be uncertain about whether an annotation contains *errors* propagated through the corpus building and annotation process, or whether the resulting annotation is error-free. Likewise, *biases* can lead to uncertainty because one might be uncertain as to how representative the corpus is, and whether certain characteristics of the data are true properties of the language, or perhaps skewed due to a particular bias. As errors and biases are often found in the data independently of any annotation scheme (see Figure 1), each can result in further annotation challenges.

5 Research Opportunities and Open Challenges

The representation problems outlined above have real consequences in terms of the reproducibility, usability and trustworthiness of research results generated via annotated resources. It is thus important to treat the sources of these problems carefully, separating them out in corpus development and use. The aim is that outlining the different sources of representation problems, and how these surface in the various phases, will lead to a more nuanced understanding of the challenges involved. This in turn can inform future resources so that they are more faithful to the data they represent, increasing the end-user's confidence with respect to research results. As an initial step, we recommend that the various types of representation problems are properly identified in linguistic annotations and labelled explicitly. In this way, we hope to be able to capture and harness the full characteristics of corpus linguistic data in the future. Our hope is that this will further lead to more robust corpus-based findings in theoretical linguistics and more accurate NLP models. Specifically, we envisage new research opportunities in theoretical and computational linguistics, as well as novel responses in connection with the reproducibility crisis.

Facilitating Theoretical Linguistic Research Modeling representation problems provides us with a clearer picture of the underlying data, furthering our understanding of the linguistic and extra-linguistic properties of the texts in a corpus. This could lead to novel research results which were previously hindered by these problems, advancing the respective state-of-the-art in theoretical linguistics. Furthermore, an explicit treatment of the problem sources could foster the emergence of new insights since, e.g., in historical linguistics, *variation* and *ambiguity* are seen as the key components of language change.

Improving NLP Models In computational linguistics, propagating representation problems throughout NLP pipelines could inform computational models at each step, improving the accuracy of the respective algorithms and the resulting end-product. For example, it has been shown that the accuracy of NLP systems for event ordering can be improved by assigning specific tags in cases of ambiguity and uncertainty (Chambers et al., 2014). Similarly, Plank et al. (2014) have shown that providing information about annotator disagreements during training of an NLP model for POS tagging increases the performance of corresponding NLP downstream models. However, more recently, Pavlick and Kwiatkowski (2019) have shown that state-of-the-art NLP models for NLI are able to model some sort of uncertainty, but this is not the uncertainty that stems from human disagreement. Therefore, they advocate the need for a better understanding of the sources of linguistic uncertainty and the downstream propagation of such

uncertainties in NLP models. Providing an NLP model with a more elaborate and explicit treatment of representation problems thus has the potential to immensely improve research results.

Promoting Reproducibility A framework which models representation problems could be of great benefit for research into the reproducibility issue in computational linguistics. Cohen et al. (2018) define reproducibility in NLP as a property related to the outcomes of an experiment with the three reproducibility dimensions *conclusion* (an induction based on the results), *finding* (relationship between values), and *value* (a calculated or measured number). These dimensions also pertain to our described problem sources, since they have an impact on whether a conclusion, a finding or a value can be reproduced. Fokkens et al. (2013) moreover show that research into the reproducibility of experiments in NLP and in particular understanding the experimental *variation* can improve research results.

Quantifying Representation Problems A future challenge will be the quantification of representation problems in linguistic annotations. For example, Likert-scales and other relational measurements have been proposed to measure the degree of uncertainty (see Section 2). We aim at experimenting with different measures in future work, exploring which measures work best for specific types of representation problems. Moreover, we intend to experiment with different kinds of probability measures in order to be able to provide a mathematical framework for propagating representation problems throughout the phases of corpus development and use.

Guided Annotation Systems To support the annotation process throughout its different phases, systems can be designed based on guidelines for identifying, capturing, and treating representation problems (Sperrle et al., 2020). Best practices from current approaches can be taken into consideration to derive such guidelines. These will establish a systematic procedure for dealing with representation problems in a consistent manner. In addition, annotation systems that rely on such guidelines could detect annotation inconsistencies and irregularities to guide annotators to the presence of representation problems, enabling them to be captured and possibly avoided or treated. Based on the users' interaction with such a learning annotation system, it can adapt over time to the users' annotation preference (Sperrle et al., 2019), cementing best practices into instructions for future sessions. Such guided systems could facilitate the annotation process and ensure the correctness of the annotation results, enabling a more reliable interpretation of the annotated data.

6 Conclusion

In this paper, we presented five sources of representation problems in linguistic annotations: ambiguity, variation, uncertainty, error and bias. We characterized these sources, outlining their usual treatment in corpus linguistic processes. Moreover, we discussed the consequences which an insufficient or improper treatment of these problems may have in the three phases of corpus development and use: data selection and processing (Phase I), annotation (Phase II) and interpretation (Phase III). In this way, this paper highlights the importance of developing more adequate and explicit treatments of such representation problems in the future. Moreover, we argued that harnessing representation problems in the scientific process fosters research into the reproducibility crisis, in addition to providing more robust research results.

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161. Many insights for the paper came from Hannah Booth's work developing the Corpus of Historical Low German ('CHLG'), with support from the Hercules Foundation/FWO, Grant number Hercules AUGE13/02 (July 2014–December 2015)/FWO G0F2614N (January 2016–present).

References

Fabian Barteld, Sarah Ihden, Ingrid Schröder, and Heike Zinsmeister. 2014. Annotating descriptively incomplete language phenomena. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 99–104, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Jana Beck, Aaron Ecay, and Anton Karl Ingason. 2015. Annotald. version 1.3. 7.

Marcel Bollmann, Florian Petran, and Stefanie Dipper. 2014. Applying rule-based normalization to different types of historical texts — an evaluation. *Human Language Technology Challenges for Computer Science and Linguistics. 5th Language and Technology Conference, LTC 2011. Revised Selected Papers*, 8387:166–177.

Georges-Pierre Bonneau, Hans-Christian Hege, Chris R. Johnson, Manuel M. Oliveira, Kristin Potter, Penny Rheingans, and Thomas Schultz. 2014. Overview and state-of-the-art of uncertainty visualization. In Charles D. Hansen, Min Chen, Christopher R. Johnson, Arie E. Kaufman, and Hans Hagen, editors, *Scientific Visualization: Uncertainty, Multifield, Biomedical, and Scalable Visualization*, pages 3–27. Springer London, London.

Hannah Booth, Anne Breitbarth, Aaron Ecay, and Melissa Farasyn. 2020. A Penn-style Treebank of Middle Low German. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 766–775, Marseille, France, May. European Language Resources Association.

Federico Boschetti, Matteo Romanello, Alison Babeu, and David Bamman. 2009. Improving OCR accuracy for classical critical editions. In *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries*, pages 156–167, 09.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Bouma, Irene Kramer, and Joost Zwarts, editors, *Cognitive Foundations of Interpretation*, pages 69–94. Royal Netherlands Academy of Science, Amsterdam.

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506, Baltimore, Maryland, June. Association for Computational Linguistics.

J. K. Chambers and Natalie Schilling. 2013. *The Handbook of Language Variation and Change*. Blackwell, Oxford, 2nd edition.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online, July. Association for Computational Linguistics.

K. Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurélie Névéol, Cyril Grouin, and Lawrence E. Hunter. 2018. Three Dimensions of Reproducibility in Natural Language Processing. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

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

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, pages 4585–4592.

Flavia De Simone, Barbara Balbi, Vincenzo Broscritto, Simona Collina, Roberto Montanari, Federico Boschetti, and Anas Kahn. 2018. The impact of human factors on digitization: An eye-tracking study of OCR proofreading strategies. In *The Tenth International Conference on Advanced Cognitive Technologies and Applications (COGNITIVE 2018)*, pages 14–17, Barcelona, Spain.

Hendrik de Smet. 2009. Analysing reanalysis. *Lingua*, 119(11):1728–1755.

Stefanie Dipper, Karin Donhauser, Thomas Klein, Sonja Linde, Stefan Müller, and Klaus-Peter Wegera. 2013. HiTS: ein Tagset für historische Sprachstufen des Deutschen. *Journal for Language Technology and Computational Linguistics*, 28:85–137.

Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. 2016. A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Jürg Fleischer. 2009. Paleographic clues to prosody? – Accents, word separation, and other phenomena in Old High German manuscripts. In Roland Hinterhölzl and Svetlana Petrova, editors, *Information structure and language change: new approaches to word order variation in Germanic*, volume 203 of *Trends in Linguistic Studies and Monographs*, pages 161–189. Mouton de Gruyter, Berlin/New York.

Antske Fokkens, Marieke van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. 2013. Offspring from reproduction problems: What replication failure teaches us. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1691–1701, Sofia, Bulgaria, August. Association for Computational Linguistics.

Jost Gippert and Ralf Gehrke. 2015. Historical corpora. Challenges and perspectives. In Jost Gippert and Ralf Gehrke, editors, *Historical Corpora. Challenges and Perspectives*, Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache 5, pages 9–12. Narr, Tübingen.

Stefan Th. Gries and Andrea L. Berez. 2017. Linguistic annotation in/for corpus linguistics. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 379–409. Springer, Dordrecht.

Bernd Heine and Tania Kuteva. 2002. *World lexicon of grammaticalization*. Cambridge University Press, Cambridge.

Bernd Heine. 2002. On the role of context in grammaticalization. In Ilse Wischer and Gabriele Diewald, editors, *New Reflections on Grammaticalization*, pages 83–101. John Benjamins, Amsterdam.

Paul J. Hopper and Elizabeth Closs Traugott. 2003. *Grammaticalization*. Cambridge University Press, Cambridge.

Paul J. Hopper. 1991. On some principles of grammaticalization. In Elizabeth Closs Traugott and Bernd Heine, editors, *Approaches to grammaticalization. Volume I. Theoretical and methodological issues*, pages 17–35. John Benjamins, Amsterdam/Philadelphia.

Eduard Hovy and Julia Lavid. 2010. Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. *International Journal of Translation*, 22(1).

Richard Hudson. 2003. Gerunds without phrase structure. *Natural Language & Linguistic Theory*, 21(3):579–615.

Markus John, Steffen Koch, and Thomas Ertl. 2017. Uncertainty in visual text analysis in the context of the digital humanities. In *Designing for Uncertainty in HCI: When does uncertainty help? (Workshop at CHI 2017)*.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall, 2nd edition.

David Jurgens. 2013. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 556–562, Atlanta, Georgia, June. Association for Computational Linguistics.

Karin Kipper Schuler. 2005. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Klaus Krippendorff. 2004. Reliability in content analysis: some common misconceptions and recommendations. *Human Communication Research*, 30(3):411–433.

Anthony Kroch. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change*, 1(3):199–244.

Merja Kytö. 2010. Corpora and historical linguistics. *Revista Brasileira de Linguıstica Aplicada*, 11:417–457, 12.

William Labov. 1994. *Principles of linguistic change. Volume I: Internal factors*. Blackwell, Oxford.

Geoffrey Leech. 1991. The state of the art of corpus linguistics. In *English Corpus Linguistics: Linguistic Studies in Honour of Jan Svartvik*, pages 8–29. Longman, London.

John Lowe. 2016. Participles, gerunds and syntactic categories. In Doug Arnold, Miriam Butt, Berthold Crysmann, Tracy Holloway King, and Stefan Müller, editors, *Proceedings of the Joint 2016 Conference on Head-driven Phrase Structure Grammar and Lexical Functional Grammar, Polish Academy of Sciences, Warsaw, Poland*, pages 401–421, Stanford, CA. CSLI Publications.

Anke Lüdeling. 2017. Variationistische Korpusstudien. In Marek Konopka and Angelika Wöllstein, editors, *Grammatische Variation. Empirische Zugänge und theoretische Modellierung. IDS Jahrbuch 2016*, pages 129–144. de Gruyter, Berlin.

Robert Malouf. 1996. A constructional approach to English verbal gerunds. *Annual Meeting of the Berkeley Linguistics Society*, 22(1):255–266.

Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. *Treebank-3*. Linguistic Data Consortium, Philadelphia.

Marie-Luis Merten and Nina Seemann. 2018. Analyzing constructional change: Linguistic annotation and sources of uncertainty. In *Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality*, TEEM'18, page 819–825, New York, NY, USA. Association for Computing Machinery.

Leonard Neidorf. 2013. Scribal errors of proper names in the Beowulf manuscript. *Anglo-Saxon England*, 42:249–269.

Rebecca J. Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694, March.

Herbert Penzl. 1967. The linguistic interpretation of scribal errors in Old High German texts. *Linguistics*, 5(32):79–82.

Susan Pintzuk. 2003. Variationist approaches to syntactic change. In Brian D. Joseph and Richard D. Janda, editors, *The Handbook of Historical Linguistics*, pages 509–528. Blackwell, Oxford.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden, April. Association for Computational Linguistics.

Randi Reppen. 2010. Building a corpus. In Anne O'Keeffe and Michael McCarthy, editors, *The Routledge Handbook of Corpus Linguistics*, pages 31–37. Routledge, London.

Beatrice Santorini. 2010. Annotation manual for the Penn Historical Corpora and the PCEEC. Department of Linguistics, University of Pennsylvania. https://www.ling.upenn.edu/hist-corpora/annotation/index.html.

Sarah Schulz and Jonas Kuhn. 2017. Multi-modular domain-tailored OCR post-correction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2716–2726, Copenhagen, Denmark, September. Association for Computational Linguistics.

Nina Seemann, Marie-Luis Merten, Michaela Geierhos, Doris Tophinke, and Eyke Hüllermeier. 2017. Annotation Challenges for Reconstructing the Structural Elaboration of Middle Low German. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 40–45, Vancouver, Canada, August. Association for Computational Linguistics.

Anders Søgaard, Barbara Plank, and Dirk Hovy. 2014. Selection bias, label bias, and bias in ground truth. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 11–13, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Fabian Sperrle, Rita Sevastjanova, Rebecca Kehlbeck, and Mennatallah El-Assady. 2019. VIANA: Visual interactive annotation of argumentation. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 11–22.

Fabian Sperrle, Mark-Matthias Zymla, Mennatallah El-Assady, Miriam Butt, and Daniel Keim. 2020. Guided linguistic annotation of argumentation through visual analytics. In *ArgVis 2020 — COMMA Workshop on Argument Visualization*.

Sali A. Tagliamonte. 2006. *Analysing sociolinguistic variation*. Cambridge University Press, Cambridge.

Siddharth Vashishtha, Benjamin Van Durme, and Aaron Steven White. 2019. Fine-grained temporal relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2906–2919, Florence, Italy, July. Association for Computational Linguistics.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 annotation manual. https://catalog.ldc.upenn.edu/docs/LDC2019T05/PDTB3-Annotation-Manual.pdf.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.