Computational text analysis within the Humanities: How to combine working practices from the contributing fields?

Jonas Kuhn¹

Published online: 26 June 2019
© The Author(s) 2019

Abstract This position paper is based on a keynote presentation at the COLING 2016 Workshop on Language Technology for Digital Humanities in Osaka, Japan. It departs from observations about working practices in Humanities disciplines following a hermeneutic tradition of text interpretation versus the method-oriented research strategies in Computational Linguistics (CL). The respective praxeological traditions are quite different. Yet more and more researchers are willing to open up towards truly transdisciplinary collaborations, trying to exploit advanced methods from CL within research that ultimately addresses questions from the traditional Humanities disciplines and the Social Sciences. The article identifies two central workflow-related issues for this type of collaborative project in the Digital Humanities (DH) and Computational Social Science: (1) a *scheduling dilemma*, which affects the point in the course of the project when specifications of the core analysis task are fixed (as early as possible from the computational perspective, but as late as possible from the Humanities perspective) and (2) the *subjectivity problem*, which concerns the degree of intersubjective stability of the target categories of analysis. CL methodology demands high inter-annotator agreement and theory-independent categories, while the categories in hermeneutic reasoning are often tied to a particular interpretive approach (viz. a *theory of literary interpretation*) and may bear a non-trivial relation to a reader's pre-understanding. Building a comprehensive methodological framework that helps overcome these issues requires considerable time and patience. The established computational methodology has to be gradually opened up to more hermeneutically oriented research questions; resources and tools for the relevant categories of analysis have to be constructed. This article does not call into question that well-targeted efforts along this path are worthwhile. Yet, it

✉ Jonas Kuhn
jonas.kuhn@ims.uni-stuttgart.de

¹ Institute for Natural Language Processing (IMS), University of Stuttgart, Stuttgart, Germany

makes the following additional programmatic point regarding directions for future research: It might be fruitful to explore—in parallel—the potential lying in DH-specific variants of the concept of rapid prototyping from Software Engineering. To get an idea of how computational analysis of some aspect of text might contribute to a hermeneutic research question, a prototypical analysis model is constructed, e.g., from related data collections and analysis categories, using transfer techniques. While the initial quality of analysis may be limited, the idea of *rapid probing* allows scholars to explore how the analysis fits in an actual workflow on the target text data and it can thus provide early feedback for the process of refining the modeling. If the rapid probing method can indeed be incorporated in a hermeneutic framework to the satisfaction of well-disposed Humanities scholars, a swifter exploration of alternative paths of analysis would become possible. This may generate considerable additional momentum for transdisciplinary integration. It is as yet too early to point to truly Humanities-oriented examples of the proposed rapid probing technique. To nevertheless make the programmatic idea more concrete, the article uses two experimental scenarios to argue how rapid probing might help addressing the scheduling dilemma and the subjectivity problem respectively. The first scenario illustrates the transfer of complex analysis pipelines across corpora; the second one addresses rapid annotation experiments targeting character mentions in literary text.

Keywords Digital humanities · Adaptation of NLP tool chains · Interdisciplinary working practice

1 Introduction

1.1 Preliminaries

Many years of research and tool development in the fields of Natural Languages Processing (NLP) and Computational Linguistics (CL) have led to (1) the availability of numerous mature tools for text analysis in the major languages, such as lemmatizers, part-of-speech taggers, parsers, etc., but also tools for specific tasks beyond linguistic annotation such as sentiment analysis, translation, purpose-specific information extraction, etc. Alongside the technical machinery (2) an advanced methodology has been developed, defining appropriate workflows for training, adapting, evaluating, and employing new analysis components where off-the-shelf tools are not readily available—because the text corpus under consideration diverges from the development standard (earlier language stage, special text genre or content domain, under-resourced language, etc.), and/or because the analytical task involves steps not covered so far (e.g., the identification of passages of scenic narration in novels and other narrative texts).

Both (1) the use of existing tools and (2) the adaptation/augmentation of analysis systems is supported by resource infrastructures such as CLARIN¹ (providing

¹ Common Language Resources and Technology Infrastructure, see www.clarin.eu, and www.clarin-d.net for the German partner within the European initiative.

access to interoperable tools and to corpora, e.g., for training data), and by publicly available code libraries. In principle, it now takes manageable effort to build or adapt text analysis systems for arbitrary combinations of text corpora and research questions (as is demonstrated by the rapid developments in Natural Language Processing over the past 5–10 years, particularly accelerated by recent successes in the application of artificial neural net models—“Deep Learning”). A number of recent contributions show that CL techniques can be expanded to construct analysis systems for literary texts. They for instance support the extraction of Social networks among literary characters (Elson et al. 2010), an analysis of the text-internal dynamics of inter-character relationships (Chaturvedi et al. 2016; Iyyer et al. 2016) or aspects of plot structure (Goyal et al. 2010), they induce types of characters from large text collections (e.g., Bamman et al. 2014) or help understand the stylistic characterization of certain character types (Brooke et al. 2017). Over the past few years, small communities of researchers pushing targeted computational modeling techniques have evolved in several field-specific branches of DH.² Yet, computational modeling components of this kind³ are still rarely used within the core areas of the classical Humanities disciplines like Literary Studies or History, which generally take a hermeneutic approach⁴ to text interpretation and, moreover, textual criticism, which is aimed at the *significance* of a text—following Hirsch’s (1967) separation of text meaning and significance, where the latter comprises the “relationship between [the text] meaning and a person, or a conception, or a situation, or indeed anything imaginable” (Hirsch 1967: 8).⁵ Under a hermeneutic approach, a literary scholar may for instance try to understand the significance of a group of novels from a particular epoch and cultural background against its historical context, possibly taking into account the sociological situation at the time etc. There may thus be multiple valid interpretations (at the level of its significance) for the same text. Often, such “polyvalence” is seen as a constitutive property of

² For Computational History, see for instance the contributions in Bozic et al. (2016); for Computational Literary Studies, many researchers came together for the 2017 International DFG Symposium in Literary Studies at Villa Vigoni, Italy 9–13 October 2017, which was dedicated to Digital Literary Studies. A volume of the contributions is in preparation (Jannidis, in preparation).

The development of the subfield of Computational Literary Studies in Germany is quite dynamic. In 2018, the Deutsche Forschungsgemeinschaft (DFG), established a priority program *Computational Literary Studies* (SPP 2207) with a runtime of 6 years.

³ This is not to say that digital editions and computational support for corpus search and exploration have not been widely adopted. Indeed, fields like Literary Studies have changed with the omnipresence of computational tools and resources, hence the observation of a “computational turn” (Berry 2011). Most influential have arguably been the contributions by Moretti (2007) and Jockers (2013). The emphasis in this article is however on the integration of formal/algorithmic models in the core argumentation of scholarly research in the Humanities disciplines, which up until today are still in its early stages.

⁴ Our focus is not on Philosophical Hermeneutics (associated with Friedrich Schleiermacher, Wilhelm Dilthey, Hans-Georg Gadamer a.o.), but on hermeneutics as a broadly adopted practical method in Literary Studies and other disciplines targeting the interpretation of works of art. As Newton (1989) points out, the term, which has long been central in the German tradition, gained prominence in the Anglo-American tradition with the work by Hirsch (1967). “The central concern of hermeneutics is the problem created by the fact that texts written in the past continue to exist while their authors and the historical context which produced them pass away in time. Reading such texts therefore becomes inseparable from the question of interpretation.” (Newton 1989: 116).

⁵ See also the discussion of the aims of text interpretation in Mantzavinos (2016).

literary texts, and it indicates that the standard CL methodology of corpus annotation and computational modeling, which aims at determining a single, intersubjectively stable target, cannot be applied at this level (we will come back to this in Sect. 5).⁶

Despite the special character of text interpretation as the final objective of hermeneutic research, much of the evidence that the scholar can build on is available in the form of preserved texts and other sources, so—from the point of view of Computational Linguists—the use of advanced corpus-based methods as a means to ensure systematicity in the process would seem an evident choice nonetheless (of course, the scholar’s pre-understanding would have to be reflected in the formulation of the corpus analysis task). On the other hand, the relevant analytical questions are likely to be different from study to study. Moreover, computational tools for the questions will rarely be readily available (few questions being directly correlated with the linguistic form of the text). Hence, it is not surprising when a specialist scholar keeps relying on their erudition and manual analysis rather than investing time into the development or refinement of analytical tools that may have just a one-time application.⁷

Given this understandably conservative tendency in the core Humanities disciplines, emerging DH fields such as Digital Literary Studies, tend to focus on questions that have not been at the center of traditional research (e.g., stylometric research⁸ or corpus-oriented research on the historical development of

⁶ An anonymous reviewer questions that there is any systematic difference between the computational modeling situation (1) for a linguistic analysis task (such as part-of-speech tagging) and (2) in the context of a hermeneutic approach to some object of study in the Humanities. He or she points out that the need to place a modeling approach within a space of competing interpretive frameworks, whose categories are mutually incompatible, is familiar from linguistic tasks: competing linguistic theories go along with different conceptual frameworks that cannot be easily reconciled in corpus-oriented operationalization. While I agree that the modeling challenges may not be categorically different between a linguistic context and for an embedding in a hermeneutic context, the two different cultural/methodological umbrellas *do* go along with substantially different weight that needs to be placed on establishing a framework of analytical concepts that (a) support an intersubjectively agreed-upon operationalization (e.g., for categories of text analysis) and at the same time (b) allow researchers to put hypotheses to a test that have been considered relevant in the recent literature. In the Humanities, the distinction between the *meaning* of a text versus its *significance* relative to various possible conceptions and situations (Hirsch 1967) discussed above goes along with a much more diverse landscape of distinct frameworks of reference, and even when text analysis is limited to the simpler notion of literal text meaning, pre-existing theoretical frameworks with a similar degree of formal specification as is common in Linguistics are rarely found in the Humanities. At the risk of oversimplifying things, one may ascribe this to modern Linguistics following a methodological agenda modeled after the natural sciences and generally working in relatively settled paradigms of “normal science” (in terms of T. Kuhn’s 1962 terminology), whereas much of the research in the Humanities does not aspire to forming a paradigm with a disciplinary consensus; the work practices there rather follow the patterns that T. Kuhn uses to characterize the “pre-paradigm” period.

Of course, it is not impossible to approach literary texts and historical source collections with an approach following the methodology established in Linguistics; but this is a different thing from trying to build on top of insights from (many distinct existing) hermeneutic branches in the established Humanities disciplines. And here, considerable cultural differences cannot be denied.

⁷ Below we will also come to the problem that even the best conceivable computational tools will display a level of accuracy that will make their use unattractive to scholars. And with tentatively small datasets underlying Humanities study, quality issues are amplified.

⁸ Viz. the success story of the metric of Burrows’s Delta (Burrows 2002), e.g., applied to German literary history by Jannidis and Lauer (2014), and scrutinized methodologically by Evert et al. (2015).

key genres⁹) rather than trying to augment the methodological spectrum for addressing classical key questions of text interpretation.

To sum up, from the CL perspective it seems that the potential of computational models for Humanities research is currently underexploited.

This article, which expands on a keynote presentation at the COLING 2016 Workshop on Language Technology for Digital Humanities (LT4DH) in Osaka, Japan,¹⁰ contributes some more detailed considerations of how this status quo can be explained and whether it could (and should) be changed. The basis for characterizing the status quo as underexploiting a potential are mostly the author's personal exchanges on many occasions—with scholars from various disciplines in the Humanities and Social Sciences. Numerous studies could in principle benefit from computational corpus analysis targeting special, non-trivial analytical categories, but given the considerable development effort and unclear chances of success (in terms of supporting innovative conclusions), it often seems wiser to follow simpler approaches.

For thoughts about potential paths along which the situation might be changed, the LT4DH keynote and this article rely mainly on experiences from collaborative DH projects involving the author himself, as well as from collaborative research involving CL and Linguistics.¹¹ This is because the emphasis here is not on research results as typically reported in publications, but more on observations about practices, gathered along the way of collaborative project work and in dedicated methodological explorations. This view makes this article a fairly subjective contribution which cannot claim to describe the status quo in a systematic way. Nor is there a claim of exclusivity. But hopefully, this contribution will stimulate further methodological discussions and developments in an exciting interdisciplinary and transdisciplinary area.

⁹ See e.g., Underwood (2015), but also earlier work such as Biber and Finegan (1989).

¹⁰ The title of the keynote presentation was *Flexible and Reliable Text Analytics in the Digital Humanities—Some Methodological Considerations*.

¹¹ The author's recent DH projects are the following: CLARIN-D (common language resource infrastructure for the Digital Humanities, 2011–2020, in which Stuttgart has been a center within the German network, led by the University of Tübingen, Erhard Hinrichs (the Stuttgart center has been co-funded by the German Federal Ministry of Education and Research BMBF and the state ministry MWK in Baden-Württemberg); the eHumanities project “e-Identity” (2012–2015, funded by the Federal Ministry BMBF, with leading PI Cathleen Kantner from Political Science and additional Co-PI's Ulrich Heid and Manfred Stede); the eHumanities project “ePoetics” (2013–2016, funded by the Federal Ministry BMBF, with leading PI Sandra Richter from Literary Studies and additional Co-PI's Thomas Ertl and Andrea Rapp); the Digital Humanities “Center for Reflected Text Analytics” CRETA (2016–2020, with a group of 10 PI's from various disciplines in the Humanities and Social Sciences, directed by the author and funded by Federal Ministry BMBF); the “DebateExplorer” project (2016–2017, a data-driven journalism project funded by Volkswagen Foundation, with Co-PI Eva Wolfangel, a journalist); the “RePlay-DH” (2016–2019, on research data management for the Digital Humanities, funded by the state ministry MWK in Baden-Württemberg, with Co-PIs Helge Steenweg and Stefan Wesner).

Collaborative project experience regarding links between CL and Linguistics include for instance involvement 2007–2010 in SFB 632 *Information structure: The linguistic means for structuring utterances, sentences and texts* (University of Potsdam and Humboldt University Berlin, funded by Deutsche Forschungsgemeinschaft, DFG) and 2010–2018 in the DFG-funded SFB 732 *Incremental Specification in Context* (University of Stuttgart, in which the author led several subprojects and was deputy director 2012–2015 and director 2015–2018).

1.2 What this article aims to achieve

This article asks what are the reasons for the observed underexploitation of advanced models from DH and CL/NLP in the Humanities and Social Sciences. A brief explanation could be that there is simply no interest in computational methods within the core disciplines—but this is clearly not the case; the fields have always been eager to adopt new approaches and in Literary Studies, for instance, the “computational turn” (Berry 2011) is considered to be in full swing. Contributions in Literary Studies make use of visualization techniques, network analysis and other methods (however not taking advantage of the full spectrum of modeling options as the computational linguist would see it).

A different explanation might be that the respective methodological prerequisites are too far apart to be reconciled. There is probably a lot to this explanation, and throughout the years there have been numerous blogs in DH forums, discussion panels and position papers observing the two cultures problem as a major obstacle.¹² But granted the cultural divide, it is surprising that it is so hard to overcome this obstacle that after many years, there is still no best-practice recipe for teams of interdisciplinary collaborators to follow.

Could it be that what Computational Linguistics has to offer in terms of deeper analytical means is generally insufficient to be integrated into hermeneutically oriented research? An anonymous reviewer expresses the suspicion that Humanities scholars are unlikely to ever accept error rates with automatic analysis tools that are significantly above human inter-annotator discrepancies. Indeed, it seems plausible that scholars would not want to move themselves into a worse starting position than when relying on “close-reading type” manual analysis. And CL tools for analysis tasks that are more complex than part-of-speech tagging do go along with considerably higher error rates, even with contemporary newswire datasets. So how would Humanities scholars ever use automatic tools for complex tasks in the text domains of their interest? There is typically much fewer training data and hence error rates are bound to be much higher. And the closer we get to interpretive questions in a hermeneutic approach, the more extreme it appears to get. With this reasoning it seems useless to seek for collaborative workflows that help modeling deeper and deeper analysis tasks: If no tool can be expected to reach acceptable error rates, one would essentially waste time. Energy seems to be better spent on improving machine learning from small datasets.

I think one should not follow this reasoning, but rather acknowledge both as important goals: machine learning from fewer data *and* interdisciplinary integration of work practices. For one thing, entirely postponing the latter would imply that after successes in the former, there would still be a very long way before the Humanities can take advantage of them. But, more importantly, I believe that even the application of analysis models with comparatively high error rates could find a reasonable home in some next-generation hermeneutic approach. Imagine for

¹² The discussion of the status of different methodologies contributing to DH is often linked with the question of definition (see e.g., the contributions in Gold and Klein 2016). Regarding the issues stemming from the combination of different cultural traditions, the discussion in Hammond et al. (2013) seems to me to be very much to the point.

instance a scholar working on a key text from some German nineteenth century author. She suspects that this text reflects influence from the author's reception of a contemporary French text (and she wants to use this to argue for some production aesthetic thesis, pointing out that the author uses certain text features to signal the intertextual references). There are some indications in diary entries that make the assumption plausible, but no certain evidence. Now, one might envisage training a computational model for intertextual links on known cases of text pairs. On indirect links, this model will have a fairly high error rate, but if it corroborates the scholar's suspicion by predicting several passages to be likely intertextual links, it does provide a valuable additional indication for argumentation (essentially following an abductive reasoning pattern). As a matter of fact, in any historic context scholars are very much used to dealing with combinations of sources with variable reliability.¹³ An effective strategy for minimizing the risk of incorrect inferences drawn from imperfect analysis components could use several strands of analysis in parallel—chosen so the error sources are likely to be independent of each other. As a result, cases of mutual agreement are very unlikely to be analytical artifacts.

It may seem a little disturbing that it is only on hypothetical grounds that we can decide whether or not one should pursue a more integrative methodology. But if it is really true that for now there are major roadblocks that prevent an effective application of deeper computational models in hermeneutic research, it would come as no surprise that there are no examples yet that show an everyday use of the idea. The present article makes the observation that there are at least two workflow-related issues that hermeneutically oriented DH projects face even when they bear a real potential for exploiting the data-driven methodology from CL: (1) a *scheduling dilemma*, which affects the point in the course of the project when specifications of the core analysis task are fixed (as early as possible from the computational perspective, but as late as possible from the Humanities perspective); (2) the *subjectivity problem*, which concerns the degree of intersubjective stability of the target categories of analysis. CL methodology demands high inter-annotator agreement and theory-independent categories, while the categories in hermeneutic reasoning are often tied to a particular interpretive approach (viz. a *theory of literary interpretation*) and may bear a non-trivial relation to a reader's pre-understanding. Building a comprehensive methodological framework that helps overcome these issues requires considerable time and patience.

The established computational methodology has to be gradually opened up to more hermeneutically oriented research questions; resources and tools for the relevant categories of analysis have to be constructed. In many cases, this includes coming up with an inventory of descriptive categories appropriate for sharing across specific research frameworks. This article does not call into question that well-targeted efforts along this path are worthwhile. Yet, it makes the following additional programmatic point: It might be fruitful to explore—in parallel—the potential lying in DH-specific variants of the rapid prototyping idea from Software Engineering. If a method of *rapid probing of analysis models* can be incorporated in a hermeneutic framework to the satisfaction of well-disposed Humanities scholars, a

¹³ Similar methodological considerations for Digital History are made in ter Braake et al. (2016).

swifter exploration of alternative paths of analysis would become possible. This may generate considerable additional momentum for transdisciplinary integration.

It is as yet too early to point to truly Humanities-oriented examples of the proposed rapid probing technique. To nevertheless make the programmatic idea more concrete, the article uses two experimental scenarios to argue how rapid probing might help addressing the scheduling dilemma and the subjectivity problem respectively. The first scenario illustrates the transfer of complex analysis pipelines across corpora; the second one addresses rapid annotation experiments targeting character mentions in literary text.

Section 2 briefly reviews the standard methodology of data-oriented model development; Sect. 3 makes some observations about the different working practices in approaches from the Humanities versus the Computational Sciences. Against this background, Sects. 4 and 5 address the scheduling dilemma and the subjectivity problem respectively, discussing ways in which they might be tackled with the idea of rapid probing. Section 6 presents a short conclusion.

2 Background

No automatic language-technological tool achieves one hundred percent correct results¹⁴—not even when it is applied to texts whose properties correspond exactly to the corpus used in tool development. And as soon as the application context deviates from the development scenario (be it due to differences in historical language stage, register, text genre, or content domain), the error rate will increase—possibly to a considerable degree, depending on circumstances (cp. Sekine 1997). By chaining up several analysis steps in a pipeline, in which each component receives as input the output of another automatic analysis, the risk of error is potentiated. Computational approaches in the Digital Humanities that address “deeper” analytical questions (e.g., questions closer to text interpretation/textual criticism in Literary Studies) are likely to employ relatively long chains of analysis¹⁵ and are thus particularly exposed to error propagation. As an additional

¹⁴ It has to be noted of course that a perfect match is not even achieved (in non-trivial analysis tasks) when the performance of two human annotators is compared. Over the past years, the fields of corpus linguistics and CL have placed considerable emphasis on the development of high methodological standards for corpus annotation work. This has resulted in various strategies for dealing with difficult annotation decisions; across the board, a clear operationalization of the decision criteria and an empirical evaluation of *inter-annotator agreement*, as an upper bound for the prediction quality that one might expect, is required. For some analysis tasks, such as part-of-speech tagging and parsing, highly developed automatic approaches have reached accuracy levels in testing that come close to the human consensus (see Giesbrecht and Evert 2009; Manning 2011).

¹⁵ This is not necessarily the case. It may be possible to employ “end-to-end” modeling techniques, which have proven very successful in recent NLP work (see e.g., Zhou and Xu 2015)—e.g., the guiding research question rests on the emotional perception of text passages from a corpus, and explicit records of such reception may be available for a set of representative texts (say, through marginal annotation in surviving text witnesses). In this case, computational models could (in principle) be trained to assign target emotions to candidate texts, without intermediate steps of linguistic analysis, such as the identification of relevant lemmas, etc. A practical problem for application of this technique in the Digital Humanities is that for robust and reliable predictions, a relatively large set of training data is needed,

issue when moving away from traditional research on the text material towards automated analysis of larger collections of source texts, the human view on each single document is eliminated from the process. This also eliminates a free “sanity check”: traditionally, sampling errors or other issues in the source selection procedure would have been noticed as a side effect of actually looking every text and applying some manual analysis.

Hence, it is not surprising when quick attempts to apply existing analysis tools to a corpus of text material taken from some Humanities scholar’s key research area lead to a certain degree of disappointment—an effect that is not uncommon in DH pilot studies: it is likely that the tools will get some obvious cases wrong (besides the unnoticed ones they get right) and whatever catches the eye as an aggregate outcome of the automated analysis will typically appear to replicate findings that are (apparently) “obvious” from scholarly research using conventional approaches. Such disappointments can have multiple reasons; an important one lies in the fact that the direct application of unmodified existing tools greatly underexploits the potential lying in the computational methodology.¹⁶ For the remainder of this article, our focus shall be on possible adjustments and extensions of the tool-based working practices in response to higher-level research questions from the Humanities.

2.1 The reference data-based methodology in Computational Linguistics

In data-oriented natural language processing (NLP), a standard methodology has been established that uses independently annotated “gold standard” data to avoid an overly impressionistic view of the usefulness of some analysis system for one’s own purposes. Indeed, before making use of any automatic model predictions it is of key importance to question the quality of a system relative to one’s corpus and the defined target analysis task: if the error rate is below a certain threshold, it may be safe to draw certain inferences despite the system being imperfect; on the other hand, if the system is unreliable for core categories of analysis, alternative approaches should be considered or problematic components should be fixed, etc.

Such an informed model application can be achieved with a conceptually simple procedure, which does however take some extra effort: whenever one plans to apply some analysis system to a new type of text data, a prior step of reference data-based quality assessment has to be performed. This means that a sample from the target-

Footnote 15 continued

which is typically unavailable for the analytical questions at hand. With small datasets, the classical pipeline approach, using intermediate steps of analysis, has the advantage that general linguistic knowledge (e.g., from lexical resources) can be taken into account to form generalization beyond the few available data points.

¹⁶ In many cases, straightforward tool application will nevertheless advance the analytical basis in corpus-oriented work and be highly beneficial (e.g., taking advantage of apparently small modifications of the view on the text, such as lemmatization or part-of-speech tagging, or through the use of unsupervised computational methods such as latent topic modeling).

specific corpus data has to be annotated manually with the intended target analysis (where the sample of reference data is representative for the application case¹⁷).

Of course, this study-specific annotation step is associated with non-negligible effort: the annotation guidelines for the original task need to be adjusted, annotators have to be trained, and a sufficiently large amount of data needs to be annotated, ideally with multiple annotations per data-point, so inter-annotator agreement measures can be taken into account (see Hovy and Lavid 2010).

When transferring existing tools and tool chains to a new task and/or target corpus, it is tempting to skip the step of reference data annotation and rather do a post hoc assessment of the system output. However, it is definitely methodologically superior to adhere to prior manual annotation of test data for evaluating quality. With a post hoc assessment of system output, there is a known bias for the analysis that is presented (Fort and Sagot 2010), so not all system errors affecting *precision* may be reliably detected—i.e., data instances that have been incorrectly assigned to the target category. System errors affecting *recall*, i.e., missing instances in the system prediction are even harder to detect without prior data annotation. The relevant instances are, by their very nature, missing in the system prediction that undergoes the post hoc assessment. Relevant cases may only be detected by chance, in case they appear in close proximity to another data instance.

To conclude, independent corpus-based evaluation following the standard methodology is a reliable way for assessing the usefulness of one or more available systems for a task at hands and for indicating where possible adjustments are needed.

2.2 System adaptation driven by reference data: Perspectives for the Digital Humanities?

The outlined quality assessment approach is not only applied in an academic context. In language-technological practice dealing with large amounts of web data, it is also quite commonly applied—at least in a rudimentary form: when a provider of language-technological web analytics is approached by a new customer who is interested in user acceptance for their products or services as reflected in web forums (maybe restaurant reviews), the provider will assemble a corpus of customer-specific development data for the task at hand. In our example this would be a sentiment analysis task, i.e., the detection and categorization of positive or negative subjective text passages in the restaurant reviews. The provider can then optimize their system for the customer, with a clear optimization objective: *precision and recall* of the automatically retrieved web documents can be maximized—possibly with a bias for one or the other depending on the analytical goals. If the new data (the target corpus) is sufficiently similar to the development data used for the established standard systems, only minor adjustments may be needed (maybe there is a dataset of hotel reviews, which turns out to be relatively

¹⁷ The representativeness of test data is critical for methodological validity. Dealing with it in a DH context is far from trivial; however, the present article cannot focus on this [some considerations are made in Kuhn (in preparation)].

similar); otherwise, it has to be decided what components need adjustment. In extreme cases, it may be necessary to rebuild all components. (Under a supervised machine learning approach, this may mean that not just a relatively small sample of test data needs to be annotated manually, but a relatively large set of training data.)

Now, approaching the technical challenges for automatic text analysis in DH and computational Social Science, it would seem natural to apply the same procedure: Given a large collection of digitized source documents (our target corpus), a representative sample is drawn as a test corpus. For this sample, the text-analytical decisions that are supposed to feed higher-level research questions are hand-annotated, following the annotation methodology from NLP (see Hovy and Lavid 2010). The test data can then drive the further system development process, similarly as sketched above.¹⁸ The approach works very well in cases where (a) the target corpus is electronically available at the beginning of the project, and (b) the text analysis steps that are needed to contribute to the main research goals are known and can be related to existing NLP tasks (e.g., named entity recognition or sentiment analysis). This is for instance a realistic scenario for extensions of the established method of Content Analysis in Social Science (cp. Krippendorff 1980), which is based on manual text annotation (or “coding”) and has always placed emphasis on a research design that can be broken down into operationalized analysis questions.¹⁹ It also works for corpus-oriented text studies that build directly on surface text properties, i.e., empirical research in Theoretical Linguistics and structuralist approaches in Literary Studies.

However, for many research scenarios from the spectrum of Humanities disciplines for which one can expect benefits from the use of computational modeling approaches, the reference data-based standard methodology cannot be straightforwardly applied: the relevant input/output relations for analysis models that may be used are not known at the beginning of the project. It is in fact one of the major project tasks to determine appropriate analytical devices informing the higher-level research question. Many scholars would point out that the hermeneutic approach they follow is in opposition to a methodologically driven preconception of the overall research agenda as a structured set of sub-questions and analysis tasks.²⁰

In order to better understand the implications of this circumstance for computational working practices, it is worthwhile clarifying the working assumptions and preferred practices of “classical” research in the Humanities versus the typical approach from CL. I will present a schematic sketch to this end in the following section.

¹⁸ In Kuhn and Reiter (2015), we argue for a working practice in text-analytically challenging digital Humanities projects which puts the annotation of reference data at the center. The Stuttgart DH center CRETA (Center for Reflected Text Analytics, www.creta.uni-stuttgart.de) implements this strategy in a transdisciplinary framework.

¹⁹ Viz. methodological developments in the e-Identity project and in CRETA (Blessing et al. 2015a, b; Overbeck 2018).

²⁰ For instance Reichert (2014) in his preface to the anthology *Big Data* observes the common concern in data-driven research that priority is given to what is methodologically feasible on the available data; Reichert speaks of an “evidence-based concentration on what is feasible with the data” (Reichert 2014: 20, translation: J.K.).

Fig. 1 Standard approach to data-driven text analysis in Computational Linguistics: (1) Left side: Texts or text elements from training data are manually annotated at levels of description that are considered relevant for analysis (possibly including intermediate levels that can serve as features for downstream pipeline components, e.g., part-of-speech tags serving as input for a syntactic parser). (2) Middle: Using appropriate model classes from machine learning and feature sets $f_1 \dots f_n$ (extracted from the text data), model parameters are estimated based on the training corpus, for instance learning to assign category labels such as c_i to the input. Alternatively, rule-based components may implement an input–output function or part of it, which is then also evaluated against the manually annotated reference data. (3) Right side: The resulting (pipeline of) models can be applied to unseen text elements, predicting an analysis according to the learned function, i.e. assigning a label from a set of possible target categories

3 Working practices in the Humanities versus Computational Linguistics

Before going into a juxtaposition of the typical research strategies, practices and workflows in the two broad scholarly fields contributing to DH, it should be noted that generalizing across “the Humanities” is certainly problematic. There is no single methodological framework across sub-disciplines in the Humanities, and even for each specific discipline, such as Literary Studies, there is a pluralism of research approaches. Yet there are commonalities in working practice clearly contrasting with the standard methodology in Computer Science and Computational Linguistics, which I here take as a basis for reflections on how insights from the two sides can best be combined.

As another proviso, note that what follows should neither be seen as a normative characterization of best practice in the disciplines, nor as an exhaustive attempt to describe working practices. It simply serves to bring out differences across fields in the *typical* approach to breaking down one’s research ideas into an agenda—this may not do full justice to the approaches, leaving commonalities across working practices aside. But for the purpose of identifying common roadblocks in transdisciplinary approaches, this should be acceptable.

Figure 1 provides a schematic characterization of the core process of developing analysis models in modern data-oriented Computational Linguistics, here showing

Fig. 2 Natural set-up for collaborative research in DH as seen from the CL angle: Humanities scholars suggest relevant corpora, help in the identification of relevant levels and categories of analysis and perform manual annotation of a subsample of the corpus which acts as reference data; computational linguists do machine learning experiments with candidate model classes, including additional tool or data resources where appropriate (e.g., additional training data that are sufficiently similar and can be included using model transfer techniques); the reference data annotated by Humanities scholars are used for the target of optimization

the decomposition of some target analysis function into two sub-modules. As noted in Sect. 2, annotated reference data play a key role in driving the project agenda.

The computational standard approach provides clear interfaces for the integration of expert knowledge about the data under consideration: gold standard annotations of the input/output relation in key modular components can be devised in close collaboration with the “domain experts”; components for which a given discipline has strong theoretical accounts can even be modeled as a rule-based system (or as a hybrid rule-based/statistical model). Consequently, the picture that computer scientists view as a natural and fruitful collaboration scheme is sketched in Fig. 2: In exchanges with the domain experts (in the DH scenario, Humanities scholars) at the beginning of a project, the requirements for text analysis components are established, and subsequently the experts develop annotation guidelines and supervise an annotation process that leads to a reliable gold standard, capturing the targeted input/output relation for computational analysis in a precise, empirically grounded way.

On this basis, the computational linguists can experiment with different algorithmic modeling approaches and optimize model parameters, so the computational system they “deliver” at the end of this process achieves the best possible quality (measured through gold standard evaluation, including the application of tests to estimate statistical significance). Of course, the development can be

by the employment of blue-hatched input–output devices available from a research infrastructure or trained on available data).

Both Figs. 2 and 4 present straightforward extensions of the respective disciplinary self understanding, and at first glance, each extension seems to capture the requirements of a DH project exhaustively while providing a natural role for the respective partner discipline. When we compare the two resulting pictures, it becomes clear however that the two types of envisaged collaborative DH project look very different. So neither Fig. 2 nor Fig. 4 can fully meet the expectations of the respective partner discipline.

The problem of the view favored from the computational angle in Fig. 2 is the following: when it is applied in a typical project cycle comprising two or three years of funding, there is a danger that the tools and models developed will not meet any real analytical need from a Humanities context: in order to allow for a thorough model development, the corpus design, specification of analytical categories and reference data annotation has to happen at an early stage. With many design decisions, it will be hard to revise them later on when the hermeneutic process approaching specific questions has revealed different analytical interests (in terms of corpus choice or analytical task). The fact that decisions on analytical targets have to happen early makes it natural to focus on relatively generic tasks and stay with readily available, well-studied corpora. This could again reinforce the impression among skeptics addressed at the beginning of Sect. 2 that computational models can at best replicate well-known results. Since the amount of available data for less studied targets of analysis will most likely be very small, they are less attractive for systematic model development.²² Finally, using target categories for analysis that are dependent on specific interpretive assumptions (which can be more helpful than generic descriptive categories in the course of hermeneutic work based on the respective pre-understanding) is not something that the strictly systematic overall approach will encourage.

When we look at the picture that seems favorable under a Humanities perspective in Fig. 4, we can make complementary observations: computational analyses are only prompted as the need arises in a hermeneutic process, hence it is desirable to allow for each course of reasoning to draw on completely different types of analysis. Also, dependence on quite specific interpretive assumptions should in principle be possible. Practically speaking however, unless the project can take indefinite time (in which case a subproject following the scheme in Fig. 2 could be triggered each time a new analysis model is required), the methodological principle of allowing appeal to some analysis procedure at any arbitrary point of the hermeneutic research process places serious limits on the depth of analysis that can be realistically performed. Entirely distinct contexts for computational analysis as suggested by the three abstract project scenarios shown in Fig. 4 will only be possible with highly generic surface-oriented tools—which means in practice that tools relying on language-specific knowledge (such as lemmatization) may already stand against methodological transfer from one scenario to the other; more corpus or task specific

²² Machine learning on small datasets goes along with a high risk of overfitting, i.e., the system will memorize the observed patterns rather than pick up systematic generalizations.

Fig. 4 Natural set-up for collaborative research in DH as seen from the Humanities angle (showing three distinct projects at the same time, each enclosed in a large partial oval): as Humanities scholars progress in their hermeneutic research process, they formulate hypotheses about a text or text corpus, which can be addressed through recourse to computational tools or models (e.g., using corpus collocation statistics to establish whether or not a key term in the text under consideration patterns with a collection of candidate texts or a background corpus). The analysis results are then incorporated in the overall argumentation and may trigger classical “close reading” steps or further steps involving computer models. Depending on the nature of the analytical step, available tools from CL or customized model solutions may be employed (drawing on additional resources); typically, even standard analysis tasks such as lemmatization and part-of-speech tagging will require tool adaptation in the DH context since the texts are not from canonical NLP domains, genres and language stages

dimensions are even more unlikely to be sharable across the scenarios. This is not only unfortunate because it underexploits the computational potential, it also implies that critical reflection of methodological implications of computational analysis cannot build on any systematic observations across contexts of application. The latter is the basis for developing principles of ‘tool criticism’, as ter Braake et al. (2016) put it.

So in short, neither of the two scenarios is a satisfactory basis when trying to take full advantage of the strengths of both sides. Certain issues affect both scenarios in the same way, especially those relating to the small size of available data for the most relevant analysis target, which will lead to *overfitting* in training and *issues of limited accuracy of machine-learned tools*.²³ As suggested in Sect. 1.1 there may be

²³ Thanks to an anonymous reviewer for pointing out that this point should not be skipped; the great advances in (neural) machine learning from the past few years mostly rely on the availability of very large amounts of training data. So, the usefulness of analytical models in many Humanities contexts will in part depend on progress made towards improved machine learning on very small datasets.

ways of embedding models with limited accuracy in a multi-strand methodology relying on abductive reasoning—if the component models match the analytical requirements. So let us leave the **small data problem** aside despite its importance from the CL angle and ask whether there could be a better synthesis of the respective working practices that helps avoid the other issues listed, among which as far as I see two types of problem are very central: I will call these the *scheduling dilemma* and the *subjectivity problem*. Section 4 is dedicated to the former, Sect. 5 to the latter.

4 The scheduling dilemma

The *scheduling dilemma* arises from the opposing principles of maximal flexibility in the content-driven choice of where to apply computational modeling components (responding to the needs arising in the hermeneutic evolution of an understanding of the textual material) versus the systematicity in the specification and decomposition of text-analytical tasks (representing a necessary basis for methodologically valid analysis models with a predictable quality). Reliable predictions in complex analysis tasks can only be achieved with a significant development effort, which requires careful planning of the analytical decision points and categories. The epistemic interest within the Humanities on the other hand can only be reasonably pursued if the procedure can react flexibly to observations which only come to attention in the course of the study through the engagement in evolving analyses of the source material.

The scheduling dilemma could in principle be solved by very generous project runtimes: whenever an open partial question arises which has some corpus-related dimension, a “proper” computational development process could be triggered (including manual corpus annotation and computational model optimization). However, in practice this is not a realistic scenario: since the usefulness of a type of computational model in a hermeneutic process is not clear until concrete analyses are available, one would often have to trigger multiple model development processes for parallel exploration, ready to discard most of them—which could create the impression of wasting valuable research time among the collaborators (who are each under publication pressure within their disciplinary home) and might lead to principled doubts about the benefit of computational models.

4.1 Approaching the dilemma with systematic bottom-up resource building

For a realistic integration of approaches, each of the two pictures from Figs. 2 and 4 have to be adjusted to the justified requirements of the partner discipline: corpus choice and annotation by Humanities scholars has to be embedded in a hermeneutic context, and vice versa there has to be room (and expert input) for systematic model development in the contexts that are deemed relevant. In other words both sides have to move (possibly moving out of their comfort zone as Hammond et al. 2013 put it).

Fig. 5 Schematic depiction of a scenario that would allow for successful synthesis of working practices: when the analysis contexts for computational models in distinct Humanities projects are sufficiently similar, computational optimization efforts can be to the benefit of more than one application case. Besides leading to better tools (most likely), this will provide richer contexts for reflecting the analytical task itself, both from a Humanities perspective and from the computational perspective

Figure 5 depicts the idea of a combined workflow, again schematically (this time limiting attention to just two distinct Humanities project contexts in the upper part).

As the gray box underlying the hermeneutic “trajectories” suggests, the complete independence of the analytical focus from considerations about computational methodological is given up: Humanities scholars commit to experimenting with computational analysis that matches a particular pattern for which machine learning model classes are partially understood and which find correspondences in other Humanities project contexts—thus generating the grounds for systematic exploration both from the technical side and from the point of view of hermeneutic integration. The computational specialists on the other hand commit to adjusting the scope of their machine learning experiments to the needs dictated by the actual context(s) of application, including choice of corpus, focus of analysis task and possibly the emphasis on theory-dependent target categories with rather limited intersubjective stability.

How can this schema be implemented in practice? Within the spectrum of possibilities there is one that requires considerable time and patience, but avoids risks with regard to potentially missing out an important component within complex analytical scenarios: building up a full pipeline or network of related subtasks in a careful bottom-up manner (necessarily making some selective choice regarding target discipline, language, genre etc.). Such a **systematic bottom-up resource building approach** is effectively what a lot of computationally oriented projects in the Digital Humanities have been taking in the past 5–10 years (see e.g., Biemann et al. 2014; ter Braake et al. 2016; Kuhn et al. 2016; Gurevych et al. 2018), with varying dynamic flexibility in the interaction between the computational side and Humanities scholars. Over time, the inventory of readily available tools is growing, such that access to reliable computational analysis models in the course of a hermeneutic process will become less and less constrained.

A disadvantage of this path (if one wants to call it a disadvantage) is that the systematic build-up of methodological insights involves extensive phases of groundwork that do not make substantial contributions to the classical core areas in the Humanities. As a consequence, there tends to be limited recognition of this groundwork. Many DH researchers however view the process as a longer-term enterprise that requires some patience. Implications for the core areas should become noticeable once the analytical machinery has been carefully formalized and modeling approaches have been adjusted to the specific needs of the field.

A risk of the systematic bottom-up approach is the following: given the loose connection with dominant research questions from the core fields, targeting text interpretation, the DH agenda may develop a momentum of its own that could push the point of convergence between computationally oriented work and the traditional fields further and further into the future. Also, it has to be noted that in many cases, the resource building for relevant subtasks cannot rely on an established inventory of descriptive categories appropriate for sharing across specific research frameworks, so the process has to be interleaved with theoretical groundwork.

4.2 An alternative strategy: rapid probing of analysis models

Without questioning the merits of the longer-term agenda of building up a more or less exhaustive pipeline of analysis models, I would here like to discuss a different strategy that would be worth while to explore in parallel. Simply put, it can be regarded as an attempt to translate the long-established concept of rapid prototyping from Software Engineering to the transdisciplinary field of computationally advanced DH, permitting for what we might call **rapid probing** of computational analysis models within a hermeneutic context. To avoid the pitfalls of the plain picture from Fig. 4, constraining principles about the choice of target models have to be assumed (essentially the gray box from Fig. 5). However, a full bottom-up regime is not necessary before assessing the usefulness of an analytical step. When a sufficiently similar model is available for rapid adaptation, relevant aspects of the behavior of the real tool (if it was built) can be anticipated. This could help make a choice among alternative candidate options, possibly saving considerable development effort in fruitless directions.

Integration of rapid probing within a truly hermeneutic approach is as yet still a programmatic idea. In such a context the assessment of prototype models, which share only certain properties of the target analysis scenario, may be harder than in typical cases of language technology development where such a strategy is more commonly applied. But if it could be made to work, the rapid probing idea has an enormous potential. As a parallel strategy besides systematic resource building, it could generate the dynamics that the patient bottom-up path tends to lack—indicating the potential that lies in deep computational analysis.

What does it take to make rapid probing work practically? The idea depends crucially on positive answers to two questions: (a) Is it possible to migrate existing complex analysis pipelines across text collections and (partial) analytical questions? To be useful, the required technical effort should be rather limited while at the same time allowing researchers to analyze a significant part of the target corpus in a robust way (though not necessarily with the highest possible quality). (b) Can the evolutionary unfolding of content-related questions in the Humanities be augmented to incorporate experiments with preliminary corpus analysis steps? These experimental analyses, along with independent analytical considerations, should help estimate the viability of expanding the preliminary model (and hence make an informed decision when alternative modeling options are available).

Neither question can be fully answered independently of the other one—a quick technical solution to (a) that does not provide appropriate starting points for critical reflection of the preliminary analysis may make addressing issue (b) essentially impossible, even for the most highly motivated team of Digital Humanities collaborators. Nevertheless focusing mainly on question (a) in this article, I would like to make the point that there can be a positive answer: for a number of non-trivial analysis tasks, analysis chains *can* indeed be ported to a different analysis task and different corpus of source material with a reasonable effort. This is particularly so for language-technological analysis tasks that build on a background pipeline of Computational Linguistics tools and uses their output representations as features for machine learning methods, which can be flexibly adjusted to study-specific content analyses. Such methods can often be “retrained” for modified target objectives (provided that the linguistic material in the target texts does not radically violate assumptions underlying the standard tools). The gray box in Fig. 5 can be seen as the axis along which rapid adaptation across projects can be performed.

4.3 Illustrating rapid probing with the “Textual Emigration Analysis” system

It is best to illustrate the abstract strategy of rapid adaptation of analysis chains with a concrete example. The web application “Textual Emigration Analysis” (TEA, Blessing and Kuhn 2014),²⁴ was designed as an example platform showcasing the exploration of biographical information using tools from Computational Linguistics.

²⁴ The platform is available at <http://clarin01.ims.uni-stuttgart.de/tea/>; the different underlying text collections discussed in this article can be selected in a pull-down menu above the world map. Blessing and Kuhn (2014) describes technical and methodological details of the approach; a tutorial for web application can be found in <http://clarin01.ims.uni-stuttgart.de/tutorial/tea.html>.

Fokkens et al. (2014) and ter Braake et al. (2016) discuss a similar system and the methodological framework it takes to integrate it in Computational History. TEA is a good example for illustrating the present methodological point since it facilitates tool chain transfer across contexts. So, the potential ways of realizing a truly Humanities-centered rapid probing scenario can be explained rather clearly with this system—even though in our examples it is still computational linguists that have experimented with the adaptations. The point of the present article remains programmatic; technical experiments are provided to make abstract methodological ideas more concrete.

TEA is based on automatic extraction of specific biographical events from large text collections, providing an interactive visualization for aggregated extraction results. Textually extracted facts provide an enormous potential for further exploration and aggregation of distributed detail information. We chose the description of a person emigrating or relocating to a different country as an appropriate test case for this platform. This type of biographical event is (a) of interest for a variety of broader analytical studies; (b) it occurs relatively frequently; and (c) it can be visualized in aggregated form geographically. There are quite a few linguistic formulations for emigration events that can be found:

- *sie übersiedelte nach Warschau*
“she relocated to Warsaw”
- *Der Weg in die Emigration [...] führte über die Schweiz und England letztlich in die USA.*²⁵
“The path to emigration [...] led through Switzerland and England, finally to the USA”
- *Später ging sie nach Norwegen, wo sie zu den prominentesten deutschen Emigranten gehörte.*²⁶
“Later she went to Norway, where she was among the most prominent German emigrants”

From a given text collection, textual descriptions of emigration from country A to country B can be extracted automatically, and the overall relation can be visualized in an interactive world map by countries of origin and destination. Figure 6 shows a screen shot of our web application with the mouse pointer over Austria. Countries that are the origin of a relocation to Austria are light red; destination countries of a relocation from Austria are light blue. A table (at the bottom) shows the absolute numbers and the relative distribution among the source and destination countries and provides hyperlinks pointing to a list of the underlying text snippets that formed the basis of the extraction.

The snippets are displayed in a pop-up window (labeled “Emigration Details”), and are again linked to the full text source. The hyperlinking makes it straightforward for users, for example, to reassure themselves that there are no errors in the automatic extraction.

²⁵ From the German Wikipedia entry for Alfred Hauptmann (1881–1948).

²⁶ German Wikipedia entry for Hanna Sandtner, née Ritter (1900–1958).

The extraction of relevant event descriptions is based on a complex analysis pipeline, starting with preprocessing of the text base, followed by a sequence of standard natural language processing (NLP) steps—tokenization, part-of-speech tagging, lemmatization, syntactic parsing—and, lastly, task-specific steps, which combine textual information with metadata or data available in (semi-) structured format, such as the country of birth of a person. The actual determination of instances of the *emigration* relation (or an emigration event)—here a three-place relation between the person, the linguistic description of the place or country of origin and the description of the destination place or country—from the various distinct linguistic realization variants, is based on supervised machine learning, i.e., the mapping is induced from a collection of manually marked training examples, taking advantage of the generalizations captured in the linguistic analyses output by the NLP tools.

As is also discussed in Blessing et al. (2015a, b) and Kuhn and Blessing (2018), a rapid adaptation of TEA's original analysis chain to other corpora and research questions is feasible (although the chain is relatively complex), and has turned out useful in practical experiments. We can distinguish a number of adaptation scenarios:

(I) *Adjusting the target relation* The target relation for which textual instances are extracted can be adjusted in an interactive training process. So, instead of emigration events the corpus can be searched for another type of event. Extraction is realized as a machine learning classifier. The *features* on which the classifier is based include the output of language-technological preprocessing tools (incl. tokenization, lemmatization, part-of-speech tagging and syntactic dependency parsing) which are run independently of the target task. The learning process can thus exploit new generalizations for the adjusted target relation exploiting the interactive training. An example relation for which we retrained the system is membership of a person in political parties and associations.²⁷

When the text material is very divergent from the typical newspaper data for which the preprocessing tools were developed, the quality of analysis degrades. However under certain circumstances the trained relations extraction component can display acceptable behavior despite faulty underlying analyses, since the machine learning may be able to compensate for systematic errors in the preprocessing output (which has no deterministic influence on the classifier decision). This means that migration of machine-learned tools can be included in exploratory prototype experiments even when the corpus material is deviant—providing some indications for the decision for or against a full adjustment. In this scenario the research team will consider the preliminary analysis results primarily from a method-oriented point of view, abstracting away from details of the content

²⁷ Of course, the aggregation steps and visualization for the emigration analysis cannot be retained for arbitrary target relations. But with a sufficiently modular design it is conceivable to include aggregation mechanisms in the prototype-oriented exploration.

In general the adjustments are however non-trivial and require close dialogue between the technical and content point of view (for discussion of the interdisciplinary approach see Kuhn 2014; Kuhn et al. 2016).

Fig. 6 Web application “Textual Emigration Analysis”: screen view after having selected of the Wikipedia-based extraction results for Austria and furthermore having activated of the detailed text instances for migration from Austria to the United States

analysis: Could higher-quality results of the same kind be useful in the process of realization? Suppressing certain aspects of sample data for some conceptual considerations is routine in Computer Science and Computational Linguistics as method-oriented disciplines. However, under a Humanities perspective such a move is highly unusual. Perhaps this point is one of the biggest hurdles for implementation of the proposed synthesis of procedural practices. And it can be seen as a central task for the recently implemented Bachelor and Master programmes in Digital Humanities to train students in this aspect of the method-oriented abstraction from certain characteristics of the research object.

In many DH projects, the text corpus to be studied is not available in fully digitized form at the project start. Even in such cases, the exploration-through-prototype-migration approach can be carried out: one or more existing corpora that are similar to the later target corpus in relevant dimensions can be used to approximate the ultimate corpus for the purposes of assessing analytical options. (Of course, the abstraction skills are strained even more in this case.)

(II) Migrating the analysis pipeline to other text sources The second type of pipeline adjustment which Blessing et al. (2015a, b) carried out for the TEA system pertains to the adjustment of a system from the original development corpus (in our case, the collection of all biographical articles in the German version of Wikipedia) to a different underlying text collection. In our experiments it was effectively possible to perform a rapid technical migration of the full pipeline to the text collection underlying the Austrian Biographical Lexicon (*Österreichisches Biographisches Lexikon*, ÖBL) in a very short time (about four

hours),²⁸ and similarly for the German Biography (*Deutsche Biographie*). This means that the aggregation and visualization functions can be readily used on different source bases. In this context, the modular design of the system architecture and the use of web services from the CLARIN-D infrastructure pays off (Mahlow et al. 2014).

The prototype is fully appropriate for an estimate of the potential argumentative benefits to be expected from a more thorough migration, for which we argue in this article. Moreover, the idea of *Linked Data* can here be exploited, i.e., the textual analysis of biographies for specific people can be juxtaposed or merged in cases where more than one collection contains an entry. This invites the exploration of discrepancies in the text sources or peculiarities in the source selection process.

Above all however, the merger of the available resources provides a valuable post-analytical framework for evaluating the tools and methods. As discussed in Sect. 2.1, the practical development of analytical models often suffers from the notorious difficulty of detecting *recall* problems. Inspecting the system output does draw attention to *precision* errors, but to detect a *recall* error, which is an omission by the system, one would have to know the set of results. And for rare phenomena, one can only approach it by putting considerable effort into random sampling of data.

If two systems are available for which one would expect the same result (at least for some of the data, e.g., a person who appears in two different biographies), a systematic comparison of the system results can be used to detect (certain types of) recall problems: For example, if system A predicts the emigration of a person X to country L, but system B (which is based on another biographical collection) does not, the origin of the discrepancy can be easily checked in the pipeline.²⁹

The following examples illustrate the comparison between the extraction of emigration events from Wikipedia and ÖBL.³⁰ In (1) and (2), the extraction results from ÖBL indicate that there were recall errors in the Wikipedia-based extraction: the construction *zog mit ihm nach England* (“moved with him to England”) in (1b), embedded in a coordination structure was not recognized; in (2b) there is a complex coordination structure too; in addition, the Wikipedia article was missing punctuation (a period after *zwangspensioniert* (“forced to retire”).

- (1) Stokes, Marianne; née Preindlsberger (*1855 in Graz; †1927 in London), painter
 - a. [ÖBL-Artikel:] *Während ihrer Stud.reisen in die Bretagne lernte sie Adrian S. kennen und übersiedelte mit ihm nach London, wo S. seitdem regelmäßig ausstellte (u. a. Fine Art Society, Grosvenor Gallery, New*

²⁸ We are grateful to Eveline Wandl-Vogt and her colleagues at the Austrian Academy of Sciences for providing the text data and sharing their expertise.

²⁹ One has to be aware however that those gaps in recall which exist in both systems (possibly for systematic reasons), stay undetected. A comprehensive evaluation of the system quality can only be done on independently sampled and hand-annotated test data.

³⁰ Typographical emphasis by the author of this article.

Gallery, Royal Acad.). [**detected**]

“During her educational journey to Bretagne, she met Adrian S. and moved with him to London, where S. since had regular expositions (a.o. F.A.S. ...)”

- b. [Wikipedia-Artikel:] *1883 lernte sie bei einem Aufenthalt in Pont-Aven in der Bretagne den englischen Maler Adrian Scott Stokes (1854–1935) kennen. Ihn heiratete sie 1884 und zog mit ihm nach England.* [**not detected**]

“1883 she met the English painter A.S.S. in Pont-Aven in Bretagne. She married him in 1884 and moved with him to England.”

- (2) Marburg, Otto (*1874 in Römerstadt; †1948 in New York City), neurologist

- a. [ÖBL-Artikel:] *1919 wurde er als Nachfolger Obersteiners Vorstand des Neurolog. Inst. 1938 emigrierte er in die USA und arbeitete als Prof. für Neurol. am College of Physicians and Surgeons der Columbia Univ., wo er ein eigenes Laboratorium hatte.* [**detected**]

“1919 he become director of the Neurological Institute, as Obersteiner’s successor. 1938 he emigrated to the USA and worked as professor for neurology at the College of Physicians and Surgeons der Columbia Univ., where he had his own laboratory.”

- b. [Wikipedia-Artikel:] *Nach dem Anschluss Österreichs 1938 wurde Marburg wie zahlreiche andere Dozenten der Wiener Universität aufgrund seiner jüdischen Herkunft zwangspensioniert Marburg und seine Frau verließen das Land und emigrierten mit Unterstützung Bernhard Sachs’ über England in die Vereinigten Staaten.* [missing punctuation: sic] [**not detected**]

“After the annexion of Austria in 1938, Marburg like many other lecturers from the university of Vienna was forced to retire due to his Jewish origin. Marburg and his wife left the country and Bernhard Sachs helped them to emigrate the the United States, via England.”

In (3)–(5) a discrepancy in extraction results indicates a precision error: in (3) a mentioned *failed* emigration attempt led to erroneous extraction in (3b) (but not in (3a)). (4) is similar in that (4a) posits a real emigration, while the biography entry refers to incorrect reports. (5b) talks about “inner emigration”, which triggered an erroneous extraction; this can be explained by a special heuristic in the analysis chain: whenever a sentence with a trigger for the emigration relation (like the noun *Emigration*) lacks information on the countries of origin and destination (or place of birth), the system falls back on the place (or country) of birth or death from the structured data. In case the countries of origin and destination are distinct, an emigration movement is postulated. In this particular case, Trient was part of Austria-Hungary in Nikodem’s youth, but our prototype system uses present-day boundaries to map place names to countries—hence the move within the country is erroneously detected as a transnational relocation. (The mentioned heuristic seems

somewhat risky, but it helps overcome substantially more precision errors than there are recall errors that it introduces. Nevertheless, the erroneous extraction of “inner emigration” cases could be overcome by a retraining of the classifier.)

- (3) Klang, Heinrich Adalbert (*1875 in Wien (Vienna); †1954 in Wien), Law scholar
 - a. [ÖBL-Artikel:] *Bemühungen um die Ausreise sowie ein Fluchtversuch nach Ungarn scheiterten.* [**correctly not extracted**]
“Requests for emigration and an attempt to escape to Hungary failed.”
 - b. [Wikipedia-Artikel:] *Mehrere Versuche legal zu emigrieren, so in die USA, Kuba und nach China, scheiterten.* [**erroneously extracted**]
“Several attempts to emigrated legally, e.g., to the USA, Cuba and China, failed.”
- (4) Kossak, Leon (*1815 in Nowy Wiśnicz; †1877 in Krakau), Polish officer and painter
 - a. [ÖBL-Artikel:] *1848 nahm K. am ung. Aufstand teil, kämpfte im Rgt. der poln. Ulanen und nahm an der Schlacht bei Világos teil. Wo er sich dann aufhielt, ist nicht bekannt, in der Literatur wird irrtümlich angegeben, er sei nach Australien ausgewandert.* [**erroneously extracted**]
“In 1848, K. participated in the Hungarian revolt, fought with the regiment of the Polish Ulans and participated in the battle near Világos. It is unknown where he lived afterwards, the literature mentions erroneously that he emigrated to Australia”
 - b. [Wikipedia-Artikel:] [**no mention of putative emigration in the text**]³¹
- (5) Nikodem, Arthur (*1870 in Trient; †1940 in Innsbruck), painter
 - a. [ÖBL-Artikel:] [**no mention of the move in the text, only travels**]
 - b. [Wikipedia-Artikel:] *Nikodem begab sich daraufhin in eine Art „innere Emigration“; nur ihm sehr Nahestehende hatten die Möglichkeit, seine Arbeiten zu sehen.* [**erroneously extracted**]
“Nikodem then went into a kind of ‘inner emigration’; only close friends and relatives had a chance to see his work.”

A third dimension of adjustment discussed in Blessing et al. (2015a, b) is system transfer across languages.

To conclude this section, we discussed ideas for overcoming a dilemma resulting from divergent scheduling priorities: A Humanities-centered view would prefer to avoid an early commitment to specific types of analytical models, while under a CL-centered view, best results are obtained with an early detailed specification of the

³¹ The Polish Wikipedia however mentions a 1-year stay in Australia, with his brother Władysław Kossak, who did emigrate there himself (according to Polish Wikipedia).

input/output relation in specific analysis steps. A standard strategy for resolving this conflict in practice is the recourse to maximally generic analysis tools, which can be applied without adaptation. This however clearly underexploits the potential lying in the development of more targeted analysis tools with knowhow from CL.

The rapid probing idea advocated here resolves the scheduling dilemma differently, aiming to encourage explorations of more complex modeling approaches to feed the hermeneutic process. A (rapid and preliminary) migration of complex analysis pipelines across structurally related subject areas can help Humanities scholars to integrate the choice of computational analysis steps in their hermeneutic process of unfolding research questions—without effectively implying a commitment to the prediction results etc. and hence not presenting a limitation, but an enrichment of the procedural practice. In the further development and fine-tuning of those computational components that seem promising for the ultimate argumentation, the rapid prototyping approach brings the advantage that details of the models can be discussed and expanded in close interaction between the Humanities scholars and the computer scientists.

5 The subjectivity problem

Orthogonal to the scheduling dilemma, an application of the data-oriented standard methodology from Computer Science/Computational Linguistics in hermeneutically oriented research contexts may run up against what one may call the ***subjectivity problem***. As laid out in Sect. 2, within the computational disciplines the “proper use” of computational modules in an analysis chain has to adhere to the established annotation-based methodology for specifying the modules’ input/output relations: annotation guidelines have to operationalize the categories of annotation, such that an intersubjectively stable observation about language use in context is captured. By measuring inter-annotator agreement in multiple annotation experiments, the effectiveness of guidelines can even be tested empirically. Target categories leading to low levels of agreement in human annotation are generally considered problematic for data-driven modeling.

Now, when aspects of literary or historical text interpretation are targeted in a text study, the postulate of intersubjectively stable “results” becomes highly controversial. In the hermeneutic context, the process of text interpretation/textual criticism (targeting the relational notion of *significance*) is not aimed at a single, “correct” target for a given text—even if the full text production context is taken into account in all facets. Rather, throughout the reception history of important texts, new interpretations have been and will be obtained, taking different points of view such as a psychological dimension, societal considerations, production aesthetics, emphasis on intertextual links with other works, etc. In most cases, a new interpretation does not invalidate earlier interpretations. In Literary Studies, a broadly shared hypothesis is that literary texts are inherently ambiguous or

“polyvalent”.³² As a consequence, for text properties connected up with interpretive differences, intersubjectively stable interpretation results cannot be assumed.

What does this imply for the applicability of the standard annotation-based methodology in the study of literary or historical texts? A plausible reaction would seem to be to completely exclude the sphere of interpretation (in the literary or historical sense) from the scope of formal/computational modeling—leaving it to traditional hermeneutics—and rather concentrate the operationalized annotation guidelines and computational modeling efforts on descriptive categories for surface-related text properties, for which intersubjective agreement can generally be reached.³³ The annotation approach in the heureCLÉA project (Gius and Jacke 2016), focusing on narrative literary texts, implements such an approach, including reconciliation steps for resolving disagreements.

At the same time, the exclusion of those text properties from formalized annotation that are contingent on interpretive decisions seems awkward too: one of the purposes of the traditional practice of (individual, subjective) text annotation has been for the reader/annotator to record one’s subjective reading impression: These may provide the basis for observing systematic patterns among text properties in a second pass. The objective of systematicity in annotation and the concession that certain annotations are influenced by subjective judgements do not necessarily exclude each other. It would seem that a computationally enhanced hermeneutic approach could benefit from computational models based on subjective annotations—even though these do not follow the rules of “proper” data-driven modeling.

5.1 Base for illustration: point of view in narrative text

The desideratum to address the subjectivity problem in more lenient ways becomes particularly clear when considering the interplay across levels of “depth” in text analysis. As I argue in Kuhn (in preparation), most categories of text analysis that would under most circumstances be considered plain descriptive—i.e., candidates for inclusion in the strict annotation methodology—can appear in ambiguous patterns, which effectively open up a disambiguation choice that depends on preference among alternative interpretations.

Consider for instance the classification of narrative point of view in narrative texts by the Austrian author Arthur Schnitzler (1862–1931). Many of his shorter narrative texts (e.g., *Berta Garlan/Frau Bertha Garlan*, 1900) are written in third-person narrative voice, limited to the subjective viewpoint of the focal character.³⁴

³² Jannidis (2003) provides a critical discussion of the polyvalence thesis; he argues against a strong version of the thesis, which claims that because of inherent polyvalence of literature it is impossible to argue—given any two interpretations of a text—that one is more appropriate than the other. Hammond et al. (2013) discuss the difference between the two cultures of Computational Linguistics vs. Literary Studies from the experience of an annotation project targeted at literary texts—they also identify the status of ambiguity as a key issue.

³³ The separation between the two distinct levels of text *meaning* and literary *significance* advocated by Hirsch (1967) could be used as a basis—although according to Jannidis et al. 2003 the assumption of an intersubjectively stable referential meaning of a text is itself controversial in the hermeneutic tradition.

³⁴ The notion of *focalization* as a differentiated narratological concept goes back to Genette (1972) (see Köppe and Kindt 2014: 208ff); numerous variants for a definition have been proposed in narratological

In his novel *The Road to the Open* (*Der Weg ins Freie*, 1908), the viewpoint of the third-person narration alternates to a certain degree between several characters' subjective viewpoints and an objective viewpoint (predominant is narration from the narrow subjective scope of the Catholic composer George von Wergenthin-Recco, but occasional passages also take the Jewish writer Heinrich Bermann's and other characters' subjective viewpoint).

At first glance, the following two passages from chapter 2 and from chapter 3 appear to be typical depictions of George's and Heinrich's viewpoint respectively.

- (6) *It was striking nine from the tower of the Church of St. Michael when George stood in front of the café. He saw Rapp the critic sitting by a window not completely covered by the curtain, with a pile of papers in front of him on the table. He had just taken his glasses off his nose and was polishing them, and the dull eyes brought a look of absolute deadness into a face that was usually so alive with clever malice. Opposite him with gestures that swept over vacancy sat Gleissner the poet in all the brilliancy of his false elegance, with a colossal black cravat in which a red stone scintillated. When George, without hearing their voices, saw the lips of these two men move, while their glances wandered to and fro, he could scarcely understand how they could stand sitting opposite each other for a quarter of an hour in that cloud of hate.*

[Arthur Schnitzler: *The Road to the Open*, translated by Horace Samuel³⁵ (Chapter 2)]

- (7) [George has just asked Heinrich a question]
Heinrich nodded. [...] He sank into meditation for a while, thrust his cycle forward with slight impatient spurts and was soon a few paces in front again. He then began to talk again about his September tour. He thought of it again with what was almost emotion. Solitude, change of scene, movement: had he not enjoyed a threefold happiness? "I can scarcely describe to you," he said, "the feeling of inner freedom which thrilled through me [...]."

George always felt a certain embarrassment whenever Heinrich became tragic. "Perhaps we might go on a bit," he said, and they jumped on to their machines.

[Arthur Schnitzler: *The Road to the Open*, translated by Horace Samuel (Chapter 3)]

Passage (6) directly and indirectly conveys sensory perceptions by George (e.g., him seeing Rapp polishing his glasses). Seemingly similar, passage (7) depicts the

Footnote 34 continued

theory. For the purposes of making the relevant interpretive distinctions in the annotation experiments I discuss below, a relatively simple characterization of psychological point of view similar to the one Wiebe (1994) is sufficient (see also Sect. 5.2 below).

³⁵ The text passages are taken from the English translation of Arthur Schnitzler's *The Road to the Open* available at <http://www.gutenberg.org/ebooks/45895>.

mental state of Heinrich, in part through direct attribution (“*He thought of it again with what was almost emotion.*”), in part through free indirect discourse (“*Solitude, change of scene, movement: had he not enjoyed a threefold happiness?*”). Annotating the subjective viewpoint accordingly would hence seem to be relatively uncontroversial (George for (6), Heinrich for (7)).

However, when looking at the novel as a whole (and at Schnitzler’s shorter, limited-viewpoint narrations), it turns out that there are many passages with an extended build-up establishing one character’s inner view, which can then be kept up for quite some time, including the perception of other character’s actions. Since formally, we find different variants of third-person narration, the transposition of whose viewpoint we are being presented is quite subtle.

Assuming that Schnitzler likes to play with this uncertainty (which is an interpretive postulate!), passage (7) can be convincingly analyzed as depicting George’s perception of Heinrich’s actions: Heinrich’s pushing of the bike is deictically related to George’s position (“*a few paces in front*”), and we do not learn about the content of Heinrich’s meditations until he begins to speak (so we can hear him through George’s ears). The most misleading sentence is “*He thought of it again with what was almost emotion.*” What appears like a switch of the narrator’s voice towards Heinrich’s inner view can of course also be free indirect discourse—conveying George’s perception of Heinrich saying “I think of it again with what is almost emotion”.³⁶ There is nothing in the passage about Heinrich’s mental state that is not conveyed through an indirect or direct quote of what Heinrich uttered in the situation. The closing sentence “*George always felt a certain embarrassment whenever Heinrich became tragic*” resolves the tension, revealing whose viewpoint we were confronted with earlier on in passage (7). (Note that this is an interpretation of the aesthetics of the passage, which presumably cannot be defended on intersubjectively uncontroversial grounds, although it could—hopefully—be made plausible by appealing to fine-grained distinctions in the linguistic form and comparisons with other passages in the novel and other texts by the author, i.e., elements of a hermeneutic process.)

So, what we can observe when analyzing text passages using largely descriptive narratological categories is the following: the inherent ambiguity of many linguistic characterizations can easily lead to situations where “deeper” interpretive decisions percolate down to more superficial ones. (In our sample scenario, an interpretive hypothesis percolates down the recursive embedding of narrative levels: are we seeing one character’s inner state or is it another character’s perception on the first one talking about his inner state?)

If one takes the subjectivity problem to exclude a formal annotation approach (because no sufficient inter-annotator agreement can be reached), then, the possibility of such percolation happening implies that there might be no level of descriptive text analysis that is perfectly “safe” from interpretive biases. Vice versa, one might take it as a plausibility argument for an approach taking certain

³⁶ For the two sentences about the September tour, the English translation makes the “transposed viewpoint” interpretation somewhat less salient than the German original: *Dann begann er wieder von seiner Septemberreise zu sprechen. Beinahe mit Ergriffenheit dachte er an sie zurück.* So one might speculate that the translator did not have the transposed interpretation in mind.

systematic modeling efforts (or formal annotations) to conditionally depend on the acceptance of some subjective pre-understanding.

5.2 Modeling subjective categorizations: Another place for rapid probing?

For the subjectivity problem, the rapid probing idea of methodological integration presented in Sect. 4.2 can also be realized based on a standard NLP analysis chain, augmented with a task-specific machine learning classifier that is trained with the rapid prototyping idea, similar as in the previous section. The corpus data and research questions are narrative literary texts, on which narratological categorizations are performed that may be correlated with interpretive decisions.

In Kuhn (in preparation), pilot experiments on a corpus of Schnitzler texts are discussed, targeting annotation of character-specific viewpoint in the narration. The idea is to explore the implications of (different subjective) interpretive pre-assumptions by integrating them in a machine learning classifier.

The experiments adopt a straightforward mention-based operationalization of point of view that is compatible with the formally precise descriptive framework worked out by Wiebe (1994) for predicting psychological point of view in narrative texts. Her model takes the form of an algorithm that predicts at each mention of a character in the linear text sequence, whether or not the previously established point of view stays the same, or whether it is shifted to the character now mentioned. The algorithm is formulated deterministically, taking into account a differentiated set of linguistic features; so whenever there are competing interpretation options, Wiebe's algorithm would enforce a decision. However, the decision relies on the auxiliary notion of *subjective elements*, which would be the natural place for including non-determinism in the algorithm.

With modern machine learning techniques, a simple mention classification framework is a sufficient basis for rapidly probing experiments testing the effects of a model that follows a particular approach to reading point of view in Schnitzler's text. Linguistic indicators (explicit attribution of speech or thought, deictic elements, adverbial modifications, reference to sensory perception, etc.) and contextual build-up, including certain patterns of character references, are included in the feature set, and so are style indicators (as Brooke et al. 2017 showed in a detailed analysis of free indirect discourse in the writings of Virginia Woolf and James Joyce).

Due to the subtle interactions, we cannot expect a machine learning approach to reliably and robustly predict the "correct" subjective viewpoint. However, following the rapid probing idea, the behavior of alternative predictive models trained on manually annotated viewpoint annotations can be systematically compared, potentially allowing for conclusions about the role different factors play; similarly, models trained on distinct texts can provide indications for a contrastive analysis.

The relevant datapoints in the machine learning classification are defined to be all mentions of characters, in their respective context. For (6), an excerpt of the above text passage, there would for instance be seven datapoints. The annotation decision is a binary decision: whether or not character referred to by the mention is the focus

of perception at the current point of narration—where what counts is the informed reading impression, i.e., readers who have the individual impression that frequent switches of viewpoint occur will make different annotations than readers who perceive long build-ups of embedded narration levels (as discussed above). For (8), the annotation would be uncontroversial: the first two mentions refer to the focus of perception, the remaining ones do not.

- (8) *[George] stood in front of the café. [He] saw [Rapp the critic] sitting by a window not completely covered by the curtain, with a pile of papers in front of [him] on the table. [He] had just taken [his] glasses off [his] nose*

Subjective annotations of this kind can be performed quite fast; for a pilot study, about a thousand data points could be annotated within a few hours. Note that no distinction between explicit thought attribution, free indirect discourse, etc., are made, since—by design—emphasis in this pilot study is placed on the pattern of switches in viewpoint.

On the dataset, supervised classifiers can be trained using a standard machine learning library (e.g., the Python library scikit-learn <http://scikit-learn.org/>). As features, the output from multiple NLP analysis tools can be used,³⁷ including syntactic structure (which is important for detecting attributions of speech and thought), co-reference, but also verb class membership (which may lead to better generalizations).

Besides training the classifier on manually annotated data, one can also experiment with systematic automatized annotations. As mentioned above, some shorter Schnitzler narrations are entirely told from a single character's viewpoint, e.g., *Berta Garlan*. Using automatic co-reference resolution, with a few minutes of manual post-correction, a dataset marking all mentions referring to the main character as “focal”³⁸ can be generated, and one can experiment with an “intertextual” model transfer: the automatized *Berta Garlan* data are used to train a supervised classifier, and this is applied to data from *Road to the Open*, in which psychological point of view varies more among the protagonists.

Table 1 displays evaluation results of a number of different training experiments on held-out test data—the idea being to give some indication of what kind of considerations can be taken. The rows in the table ((A) through (C)) vary the training and test data in the experiments, and in the columns, evaluation results for classifiers trained with Logistic Regression are shown.³⁹

³⁷ The experiments, which were run on the English translation of the Schnitzler texts, were based on the output of the Stanford CoreNLP tools (<https://stanfordnlp.github.io/CoreNLP/>). In addition, the verb classification from VerbNet (<https://verbs.colorado.edu/verb-index/>), and the stylistic profile vectors from Brooke et al. (2017) were used, which are included in the Gutentag tools suite (www.cs.toronto.edu/~jbrooke/gutentag/); the style profile lexicon is available as `built_in_lexicons/sixstyleplus.txt`; thanks to Julian Brooke).

³⁸ What we call “focal” here is operationalized as evoking the reader's interpretation of the character being the present exponent of subjective point of view.

³⁹ F-Score is the harmonic average of precision and recall.

In scenario (A), a classifier is trained on *Road to the Open* and tested on manually annotated test data from *Berta Garlan*. The fact that relatively decent accuracy scores (0.78) can be reached in transfer across texts seems to indicate that the model picks up a certain level of abstraction (across texts, it cannot be highly text-specific clues that help in testing).

In scenario (B), with training and test data from the same text (but with a smaller training set than in (A)), the prediction accuracy is slightly lower than in (A) (0.75).⁴⁰ Scenario (C) includes “mixed” training data, testing on the same data as in (B). The classifier benefits from the increased amount of training data—which could be an indication for relatively homogeneous patterns of narrative viewpoint across texts.

As a last thing, it can be interesting for a text study aiming at interpretive aspects to check the machine-learned classifier on other texts or parts of the development text that were not taken into account in the annotation. About *Road to the Open* it has for example been observed that George’s mistress, Anna Rosner, is very rarely focalized. One can now look for text passages in which the classifier (trained on viewpoint contexts for other characters) nevertheless predicts a subjective viewpoint for clusters of references to Anna. Passage (9) (from the end of chapter 2) is an example of such a passage. This can be compared with passages where references to her are assigned low scores by the subjective viewpoint classifier, e.g., (10) (from the beginning of chapter 3).

- (9) *She had for the first time in her life the infallible feeling that there was a man in the world who could do anything he liked with her.*
- (10) *Anna had given herself to him without indicating by a word, a look or gesture that so far as she was concerned, what was practically a new chapter in her life was now beginning.*

(9) is indeed one of the few passages in which the narrator merges with Anna’s subjective perception, whereas in (10), the subjective viewpoint is intuitively George’s. So, we do find indications that a machine-learned classifiers, which a scholar can adjust to his or her individual pre-understanding within a few hours, could indeed be of use for advanced text-analytical explorations.

Of course, if a pilot study converges on certain correlations, structural patterns, etc., tentative insights from rapid probing have to be followed up by efforts for building relevant components more systematically and subjecting the models to a strict empirical evaluation on independently obtained target annotations.

⁴⁰ In experiments not shown in the table, we trained a different classification algorithm, Gaussian Naive Bayes, which generally led to lower accuracy. However it benefitted more from scenario (B); here, scenario (A) resulted in an accuracy of 0.67, compared to 0.74 for (B). Comparisons of this kind, plus error analyses etc., may help to develop a better understanding of the analysis problem.

Table 1 Experimental results from pilot study on narrative viewpoint classification

		Logistic regression classifier			
		Precision	Recall	F-score	Accuracy
(A) Training: <i>Road</i> , Test: <i>Garlan</i>	foc.	0.77	0.71	0.74	0.78
	non-foc.	0.79	0.84	0.82	
(B) Training: <i>Road</i> , Test: <i>Road</i>	foc.	0.73	0.77	0.75	0.75
	non-foc.	0.76	0.73	0.74	
(C) Training: <i>Road</i> + <i>Garlan</i> , Test: <i>Road</i>	foc.	0.76	0.73	0.75	0.79
	non-foc.	0.81	0.83	0.82	

6 Conclusion

This position paper took as its point of departure the observation that in research targeted at the literary or historic interpretation or *significance* of texts in a broadly hermeneutic sense, the use of complex computational modeling components is still an exception—despite substantial advances in the development of highly adaptable computational frameworks and infrastructures supporting re-use of tools, corpus resources and annotations. I argued that this can be explained (at least in part) through diverging working practices in humanities disciplines (predominantly following a hermeneutic tradition of text interpretation) versus the method-oriented research strategies in Computational Linguistics (with a relatively strict conception of the “proper use” of text analysis models). The diverging methodological principles pose a challenge for a joint methodology/working practice. Specifically, we can identify a scheduling dilemma that makes it hard to deploy sophisticated computational analysis chains in specialized hermeneutic studies, and the subjectivity problem. The latter originates from the constraint that in standard data-driven modeling, gold-standard annotations have to be grounded in operationalized categories leading to high inter-annotator agreement. But hermeneutic text analysis may explore implications of pre-understandings of a subjective kind, grounded for instance in aesthetic judgements.

There are ongoing efforts of systematically extending the basis for computational models targeted at research questions from the Humanities that avoid methodological clashes, which were not in the focus of this contribution. I conjecture that moving along this “main” path of systematic bottom-up resource building will expand the established computational methodology to gradually open up to more hermeneutically oriented research questions. Yet, as a swifter way of exploring the potential for methodological innovations, it may be fruitful to occasionally run rapid probing experiments, in which the established constraints on the “proper use” of computational models are consciously weakened in order to facilitate experimental investigations—also in scenarios for which no carefully developed datasets and tool chains are in place, or where researchers would like to explore implications of their own or another scholar’s subjective pre-understanding. Ideally, the increased

flexibility will inspire experimentation in more risky interdisciplinary terrain, and the successes (and failures) help assess the value of novel ideas that are otherwise too far out of the established methodological reach.

Advocating a rapid probing approach of course bears a certain risk. Swiftly achieving preliminary analytical results on a text or corpus of interest may tempt researchers to jump to conclusions. A rapidly transferred prototype model may have flaws that make the text analysis (incorrectly) appear to corroborate a scholar's pre-understanding regarding an important hypothesis (where the pre-understanding may or may not be erroneous). It is hence crucial to apply the rapid probing approach under a regime that follows highest methodological standards, which means that no substantial conclusions may be drawn from a tool's prediction unless it has been evaluated against independently annotated test data in the target domain.

What we have here is a small-scale version of the re-occurring methodological dispute: should it only be the firmly established methods that drive the research agenda—methods from the core of the research paradigm for which the community has a clear consensus (potentially at the cost of missing out on questions that cannot be fully stated within the framework)? Or is it legitimate to pursue more unorthodox research ideas exploring the borderlines of the established consensus (without sacrificing crucial standards of validity)? Rapid probing with careful subsequent empirical evaluation seems a viable instrument for proponents of the latter self-understanding. And I believe that with the current emphasis on heavily data-driven approaches to language and text, it can be healthy and fruitful to encourage such alternative practices. Indeed, a (broadly) hermeneutic approach differs from the mainstream paradigm in that the research process typically starts out from the object of study—a text or text collection—and generates questions by taking contextualizing views of the object of study (not applying filters of what are feasible methodological ways for addressing the questions until later in the process). It will still be only possible to address certain questions on a computational basis; but an approach that subordinates choice of methods to the collection of questions prompted by contextualizing the object of study will generate a different awareness of biases and systematic gaps.

Ultimately, the established (continuously growing) catalogue of NLP tasks that can be “solved” with measurable quality bears its own risk: it is easy to be overly optimistic about the degree of scholarly and scientific understanding that trainable predictive models give us about language, culture and cognition. Here too, alternative paths in research practice may help overcome biases and unveil gaps.

Acknowledgements The author would like to thank the organizers of the COLING 2016 Workshop on Language Technology for Digital Humanities (LT4DH) in Osaka, Japan, for the opportunity to engage in timely discussions about possible contributions that Language Technology and Computational Linguistics can make to the Digital Humanities. Many thanks go to five anonymous reviewers for very valuable comments and suggestions on manuscript versions of this article and to the colleagues in Stuttgart for contributions, discussions and comments, in particular to André Blessing, Nils Reiter and Gabriella Lapesa. This work was partially supported through a grant from the German Federal Ministry of Education and Research (BMBF) for the Center for Reflected Text Analytics (CRETA). Parts of Sects. 3 and 4 are based on methodological discussions in Kuhn and Blessing 2018 (contributed by the first author); an English translation is used with friendly permission by the editors of the volume “Europa baut

auf Biographien. Aspekte, Bausteine, Normen und Standards für eine europäische Biographik” (Bernád et al. 2018).

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Bamman, D., Underwood, T., & Smith, N. A. (2014). A Bayesian mixed effects model of literary character. In *Proceedings of the 52nd annual meeting of the Association for Computational linguistics*, Baltimore (pp. 370–379).
- Bernád, Á. Z., Gruber, C., & Kaiser, M. (Eds.). (2018). *Europa baut auf Biographien: Aspekte, Bausteine, Normen und Standards für eine europäische Biographik*. Vienna: New Academic Press.
- Berry, D. M. (2011). The computational turn: Thinking about the Digital Humanities. *Culture Machine*, 12, 1–22.
- Biber, D., & Finegan, E. (1989). Drift and the evolution of English style. A history of three genres. *Language*, 65, 487–517.
- Biemann, C., Crane, G. R., Fellbaum, C. D., & Mehler, A. (Eds.). (2014). *Computational Humanities—Bridging the gap between computer science and Digital Humanities*. Report from Dagstuhl Seminar 14301.
- Blessing, A., Glaser, A., & Kuhn, J. (2015b). Biographical data exploration as a testbed for a multi-view, multimethod approach in the Digital Humanities. In S. ter Braake et al. (Eds.), *Proceedings of the 1st conference on biographical data in a digital world* (pp. 53–60). Amsterdam.
- Blessing, A., Kliche, F., Heid, U., Kantner, C., & Kuhn, J. (2015a). Computerlinguistische Werkzeuge zur Erschließung und Exploration großer Textsammlungen aus der Perspektive fachspezifischer Theorien. In C. Baum & T. Stäcker (Eds.), *Grenzen und Möglichkeiten der Digital Humanities*. (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1).
- Blessing, A., & Kuhn, J. (2014). Textual emigration analysis (TEA). In *Proceedings of the 9th international conference on language resources and evaluation (LREC'14)* (pp. 2089–2093). Reykjavik: European Language Resources Association (ELRA).
- Bozic, B., Mendel-Gleason, G., Debruyne, C., O'Sullivan, D. (Eds.). (2016). *Computational history and data-driven Humanities: Second IFIP WG 12.7 international workshop*. Springer.
- Brooke, J., Hammond, A., & Hirst, G. (2017). Using models of lexical style to quantify free indirect discourse in modernist fiction. *Digital Scholarship in the Humanities*, 32, 234–250.
- Burrows, J. (2002). ‘Delta’: A measure of stylistic difference and a guide to likely authorship. *Literary and linguistic Computing*, 17, 267–287.
- Chaturvedi, S., Srivastava, S., Daumé III, H., & Dyer, C. (2016). Modeling evolving relationships between characters in literary novels. In *Proceedings of the 30th AAAI conference on artificial intelligence* (pp. 2704–2710).
- Elson, D. K., Dames, N., & McKeown, K. R. (2010). Extracting social networks from literary fiction. In *Proceedings of the 48th annual meeting of the Association for Computational linguistics, ACL'10* (pp. 138–147). Stroudsburg, PA: Association for Computational Linguistics.
- Evert, S., Proisl, T., Jannidis, F., Pielström, S., Schöch, C., & Vitt, T. (2015). Towards a better understanding of Burrows’s Delta in literary authorship attribution. In *Proceedings of the 4th workshop on computational linguistics for literature* (pp. 79–88). Denver.
- Fokkens, A., ter Braake, S., Ockeloen, N., Vossen, P., Legêne, S., & Schreiber, G. (2014). BiographyNet: Methodological issues when NLP supports historical research. In *Proceedings of the 9th edition of the language resources and evaluation conference (LREC)* (pp. 3728–3725). Reykjavik.
- Fort, K., & Sagot, B. (2010). Influence of pre-annotation on POS-tagged corpus development. In *Proceedings of the 4th linguistic annotation workshop, ACL 2010* (pp. 56–63). Uppsala.
- Genette, G. (1972). *Figures III*. Paris: Seuil.

- Giesbrecht, E., & Evert, S. (2009). Is part-of-speech tagging a solved task? An evaluation of POS taggers for the German web as corpus. In *Proceedings of the 5th web as corpus workshop* (pp. 27–35).
- Gius, E., & Jacke, J. (2016). Zur Annotation narratologischer Kategorien der Zeit. Guidelines zur Nutzung des CATMA-Tagsets. Version 2, Hamburg.
- Gold, M. K., & Klein, L. F. (Eds.). (2016). *Debates the Digital Humanities*. Minneapolis: University of Minnesota Press.
- Goyal, A., Riloff, E., & Daumé III, H. (2010). Automatically producing plot unit representations for narrative text. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 77–86).
- Gurevych, I., Meyer, C. M., Binnig, C., Fürnkranz, J., Kersting, K., Roth, S., & Simpson, E. (2018). Interactive data analytics for the Humanities. In A. Gelbukh (Ed.), *Interactive data analytics for the Humanities. Computational linguistics and intelligent text processing: Proceedings of the 18th international conference*. Lecture Notes in Computer Science (Vol. 10761). Berlin: Springer.
- Hammond, A., Brooke, J., & Hirst, G. (2013). A tale of two cultures: Bringing literary analysis and computational linguistics together. In *Proceedings of the NAACL 13 workshop on computational linguistics for literature* (pp. 1–8). Atlanta, GA.
- Hirsch, E. D., Jr. (1967). *Validity in interpretation*. New Haven: Yale.
- Hovy, E., & Lavid, J. (2010). Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. *International Journal of Translation*, 22(1), 13–36.
- Iyyer, M., Guha, A., Chaturvedi, S., Boyd-Graber, J., & Daumé III, H. (2016). Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of NAACL-HLT 2016* (pp. 1534–1544).
- Jannidis, F. (2003). Polyvalenz—Konvention—Autonomie. In F. Jannidis, G. Lauer, M. Martínez, & S. Winko (Eds.), *Regeln der Bedeutung. Zur Theorie der Bedeutung literarischer Texte* (pp. 3–30). Berlin: de Gruyter.
- Jannidis, F. (Ed.). *Digitale Literaturwissenschaft*. Stuttgart: Metzler. (**in preparation**).
- Jannidis, F., & Lauer, G. (2014). Burrows's Delta and its use in German literary history. In M. Erlin & L. Tatlock (Eds.), *Distant readings. Topologies of German culture in the long nineteenth century* (pp. 29–54). Rochester.
- Jannidis, F., Lauer, G., Martínez, M., & Winko, S. (2003). Der Bedeutungsbegriff in der Literaturwissenschaft. Eine historische und systematische Skizze. In F. Jannidis, G. Lauer, M. Martínez, & S. Winko (Eds.), *Regeln der Bedeutung. Zur Theorie der Bedeutung literarischer Texte* (pp. 3–30). Berlin: de Gruyter.
- Jockers, M. L. (2013). *Macroanalysis. Digital media and literary history*. Illinois: University of Illinois Press.
- Köppe, T., & Kindt, T. (2014). *Erzähltheorie*. Reclam: Eine Einführung.
- Krippendorff, K. (1980). *Content analysis; An introduction to its methodology*. Beverly Hills, CA: Sage.
- Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.
- Kuhn, J. (2014). The challenge for the computational sciences in Digital Humanities: Establishing a common meta-methodological framework. In *Book of abstracts zum DHd workshop "Informatik und die Digital Humanities"*. Leipzig. http://dhd-wp.hab.de/files/book_of_abstracts.pdf. Accessed 9 May 2019.
- Kuhn, J. (2017). Empirie—Beschreibung—Interpretation: über den Platz von Computermodellen in den hermeneutisch-historisch orientierten Literaturwissenschaften. In F. Jannidis (Ed.), *Article based on contribution to the international symposium of the Deutsche Forschungsgemeinschaft (DFG) on digital literary studies*, 9–13 October 2017, Villa Vigoni, Italy. (**in preparation**).
- Kuhn, J., Alexiadou, A., Braun, M., Ertl, T., Holtz, S., Kantner, C., Misselhorn, C., Padó, S., Richter, S., Stein, A., & Zittel, C. (2016). CRETA (Centrum für reflektierte Textanalyse)—Fachübergreifende Methodenentwicklung in den Digital Humanities. In *Konferenzabstracts zur Konferenz Digital Humanities im deutschsprachigen Raum* (pp. 340–343). Leipzig. <http://dhd2016.de/boa.pdf>. Accessed 9 May 2019.
- Kuhn, J., & Blessing, A. (2018). Die Exploration biographischer Textsammlungen mit computerlinguistischen Werkzeugen—methodische Überlegungen zur Übertragung komplexer Analyseketten in den Digital Humanities. In A. Z. Bernád, C. Gruber & M. Kaiser (Eds.), *Europa baut auf Biographien: Aspekte, Bausteine, Normen und Standards für eine europäische Biographik* (pp. 225–258). Vienna: New Academic Press.

- Kuhn, J., & Reiter, N. (2015). A plea for a method-driven agenda in the Digital Humanities. In *Global Digital Humanities*. Sydney.
- Mahlow, C., Eckart, K., Stegmann, J., Blessing, A., Thiele, G., Gärtner, M., & Kuhn, J. (2014). Resources, tools, and applications at the CLARIN center Stuttgart. In *Proceedings of the 12th edition of the KONVENS conference* (pp. 11–21). Universitätsverlag Hildesheim.
- Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In A. Gelbukh (Ed.), *Computational Linguistics and intelligent text processing, 12th international conference, CICLing 2011, proceedings, part I. Lecture notes in computer science* (Vol. 6608; pp. 171–189). Berlin: Springer.
- Mantzavinos, C. (2016). Hermeneutics. In E. N. Zalta (Ed.) *The Stanford encyclopedia of philosophy* (Winter 2016 edn.). <https://plato.stanford.edu/archives/win2016/entries/hermeneutics/>. Accessed 9 May 2019.
- Moretti, F. (2007). *Graphs, maps, trees: Abstract models for a literary history*. London: Verso.
- Newton, K. M. (1989). Hermeneutics and modern literary criticism. *The British Journal of Aesthetics*, 29(2), 116–127.
- Overbeck, M. (2018). Vom Beobachter zum Teilnehmer. In *Das Narrativ von der Wiederkehr der Religion* (pp. 231–260). Wiesbaden: Springer.
- Quan-Haase, A., & Martin, K. (2012). The continuing role of serendipity in historical research. In *Proceedings of the annual conference of CAIS*.
- Reichert, R. (Ed.). (2014). *Big data. Analysen zum digitalen Wandel von Wissen, Macht und Ökonomie*. Bielefeld: Transcript.
- Sekine, S. (1997). The domain dependence of parsing. In *Proceedings of the 5th conference on applied natural language processing* (pp. 96–102). Association for Computational Linguistics.
- ter Braake, S., Fokkens, A. S., Ockeloen, C. J., & van Son, C. M. (2016). Digital History: Towards New Methodologies. In B. Bozic, G. Mendel-Gleason, C. Debruyne & D. O'Sullivan (Eds.), *Computational history and data-driven Humanities: Second IFIP WG 12.7 international workshop, CHDDH 2016, Dublin, Ireland, May 25, 2016, revised selected papers. IFIP Advances in Information and Communication Technology* (Vol. 482, pp. 23–32). Springer.
- Underwood, T. (2015). Understanding genre in a collection of a million volumes. *NEH White Papers*.
- Wiebe, J. M. (1994). Tracking point of view in narrative. *Computational Linguistics*, 20(2), 233–287.
- Zhou, J., & Xu, W. (2015). End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the 53rd annual meeting of the Association for Computational linguistics and the 7th international joint conference on natural language processing* (pp. 1127–1137).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.