Ein neues Format für die Digital Humanities: Shared Tasks. Zur Annotation narrativer Ebenen

Willand, Marcus
marcus.willand@gs.uni-heidelberg.de
Universität Heidelberg; Universität Stuttgart

Gius, Evelyn
evelyn.gius@uni-hamburg.de
Universität Hamburg

Reiter, Nils
nils.reiter@ims.uni-stuttgart.de
Universität Stuttgart

Einleitung

Dieses Paper führt unsere letztjährige Präsentation[1] des Vorhabens fort, den ersten **Shared Task (ST) zur Annotation literarischer Phänomene** zu organisieren und solch ein kompetitives Verfahren als fruchtbares Format für die *Digital Humanities* einzuführen. Dieser ST hat mit dem abgehaltenen Workshop der teilnehmenden Teams einen Meilenstein erreicht.

Bei einem ST bewerben sich Teams mit einem Vorschlag für die Lösung eines durch die Organisatoren ausgeschriebenen Problems, den Task. STs sind kompetitive Verfahren, weil die Lösungsvorschläge vergleichend evaluiert und gemäß einer definierten Metrik in eine Rangfolge gebracht werden. Vor allem in der Sprachverarbeitung (NLP, natural language processing) sind diese Arbeitszusammenhänge weit verbreitet und ein wesentlicher Antrieb für die Fortschritte bei wichtigen Aufgaben, etwa des syntaktischen Parsings.[2] Wir haben dieses kompetitive Verfahren für literaturwissenschaftliche Problemstellungen durch kooperative Aspekte modifiziert und möchten hier sowohl den angepassten Workflow als auch zentrale Einsichten vorstellen, die wir durch den o.g. Workshop generieren konnten.[3] Wir gehen davon aus, dass durch solch adaptierte STs sehr viele andere Problemstellungen der Geisteswissenschaften adressiert werden können, wodurch sich STs als Verfahren für die Digital Humanities natürlicherweise anbieten. Dies ist insbesondere der Fall, wenn computationelle Verfahren auf geisteswissenschaftliche Konzepte treffen und diese in einem intersubjektiven Aushandlungsprozess operationalisiert werden sollen.

Wir haben uns für ein zweiphasiges Verfahren entschieden. Die erste Phase – „SANTA“ genannt: Systematic Analysis of Narrative Texts through Annotation – widmet sich der Erstellung von Annotationsrichtlinien für das Phänomen narrativer Ebenen.[4] Die von den acht teilnehmenden Teams eingereichten und auf dem Workshop diskutierten Richtlinien bilden die Grundlage für den Task der geplanten zweiten Phase: die automatisierte Identifikation von Erzählebenen auf Basis von Daten, die nach den Richtlinien annotiert wurden (wird vsl. 2019 ausgeschrieben).[5]

Die acht Teams divergieren hinsichtlich ihrer:

- **Größe**: 1-4 Mitglieder, wobei drei Teams die Richtlinien im Seminarkontext, also mit einer Vielzahl an Studierenden entwickelten (was eine von uns vorgeschlagene Option war)
- **Disziplin**: Literaturwissenschaft, Informatik, Linguistik, Computerlinguistik, Mediävistik, *Digital Humanities*
- **Forschungsziele**: Narratologische Konzeptentwicklung, bzw. -reflexion, Anwendung narratologischer Konzepte für die Einzeltextinterpretation (bzw. ausschließlich für die Texte im von uns vorgegebenen Korpus), linguistische Diskursanalyse, Automatisierung der Annotation
- **Nation**: Deutschland, USA, Schweden, Irland, Kanada
- **Narratologie**: Überwiegend Genette und/oder Ryan, teilweise linguistische oder selbstentworfene Level-Definitionen

Der Workshop selbst war konzeptionell offen angelegt und sollte den Teilnehmer/innen wie auch uns Organisator/innen ermöglichen, den geplanten Ablauf in Reaktion auf die Arbeitsergebnisse zu verändern. Dies war realisierbar, da bis auf wenige Kurzvorträge (z.B. stellte jedes Team zu Beginn in 5 Minuten die zentralen Aspekte seiner Richtlinien vor) hauptsächlich in kooperativen Formaten wie Gruppenarbeiten, Feedback-Runden und Plenumsdiskussionen gearbeitet wurde.

Guidelines: Unterschiede und Gemeinsamkeiten

Am **ersten Workshop-Tag** sollten die Teilnehmer/innen einen differenzierten Einblick in alle acht Annotationsrichtlinien und deren Spezifika bekommen. Dabei wurden **Unterschiede** auf mehreren Abstraktionsebenen identifiziert: Die erste und grundlegendste Einsicht bestand in der Beobachtung, dass die Definitionen narrativer Level mitunter stark differieren und diese Unterschiede durch die divergierenden Forschungsfragen (siehe oben) erklärt werden können: etwa, ob die Level-Annotation im Dienste narratologischer Konzeptentwicklung eingesetzt wird oder zur Erkennung linguistischer Diskursebenen in Texten. Das spezifische Level-Verständnis hat gleichsam Auswirkungen auf den

Modus des Definierens. So werden narrative Level teilweise inhaltlich bestimmt über die Elemente der „Story“ oder aber – etwas abstrakter – über die Elemente der Erzählung der Story. Zu unterscheiden sind davon Ansätze, die narrative Level über ihre Grenzen bestimmen, teilweise ohne auf die Erzählinhalte zurückzugreifen. Dies ist der Fall, wenn Erzähler- oder Weltwechsel als Definiens eingesetzt werden. Deutlich wurde zudem, dass gerade literaturwissenschaftliche Ansätze nicht immer eindeutig zwischen *identifizierenden* und (bloß) *charakterisierenden* Texteigenschaften narrativer Ebenen – wie etwa Fokalisierung – unterscheiden, wobei die Frage aufkommt, ob letztere einen berechtigten Ort in den Guidelines haben. **Gemeinsamkeiten** der Guidelines wurden ebenfalls diskutiert. Dabei zeigte sich, dass die eingereichten Guidelines zwei recht homogene Gruppen bilden hinsichtlich *Forschungsziel* (Narratologie vs. Automatisierung) und *Konzeptverständnis* (komplex vs. vereinfachend).

Evaluation der Guidelines: Drei Bewertungsdimensionen

Der **zweite Workshop-Tag** widmete sich der Evaluation der Guidelines: Das Ziel der Organisator/innen war es, mit den Teilnehmer/innen gemeinsam die ‚besten‘ Richtlinien zu finden, wobei zunächst unklar blieb, ob es *einen* oder *mehrere* Gewinner geben sollte (etwa jeweils einen aus den beiden o.g. recht homogenen Gruppen). Die von den Organisator/innen im Vorfeld ausgearbeiteten Evaluationskriterien zur Beurteilung der Stärken und Schwächen der Einreichungen folgen der Idee, den interdisziplinären Aushandlungsprozess anhand von drei Dimensionen zu strukturieren und so neben in der Computerlinguistik etablierten Kriterien weitere relevante geisteswissenschaftliche Aspekte in die Bewertung zu integrieren.

Diese Kriterien wurden zuerst während des Workshops im Plenum reflektiert und anschließend per online-Fragebogen live in die Evaluation überführt. Jede Dimension sollte auf nachvollziehbare Weise potentielle Guideline-Stärken vergleichbar machen und so für eine ausgewogene Beurteilung durch die Workshopteilnehmer/innen sorgen. Die drei Dimensionen sind – um in der Tagungssprache zu bleiben – *Conceptual Coverage*, *Applicability* und *Usefulness*.

Die erste Dimension beurteilt anhand von vier Fragen die Qualität der Guidelines hinsichtlich ihrer **Abdeckung der zugrundeliegenden (meist narratologischen) Theorie**:

1. Is the narrative level concept explicitly described?
2. Is the narrative level concept based on existing concepts?
3. How comprehensive are the guidelines with respect to aspects of the theory?
4. How adequate is the narrative level concept implemented by this guidelines in respect to narrative levels?

Die zweite Dimension evaluiert die **Anwendbarkeit der Guidelines auf den Text** anhand von zwei Fragen und eines von uns im Vorhinein gemessenen *Inter-annotator agreements*:[6]

1. How easy is it to apply the guidelines for researchers *with* a narratological background?
2. How easy is it to apply the guidelines for researchers *without* a narratological background?

Die dritte Dimension bewertet anhand von vier Fragen, wie der auf Basis einer Richtlinie annotierte Text das T **extverstehen und die weitergehende Textarbeit** befördert:

1. Thought experiment: Assuming that the narrative levels defined in the annotation guidelines can be detected automatically on a huge corpus. How helpful are these narrative levels for an interesting corpus analysis?
2. How helpful are they as an input layer for subsequent corpus or single text analysis steps (that depend on narrative levels)?
3. Do you gain new insights about narrative levels in texts by applying the foreign guidelines, compared to the application of your own guidelines?
4. Does the application of these guidelines influence your interpretation of a text?

Jede der Fragen wurde von den Teams in einer Feedback-Runde erläutert und anhand einer vierstufigen Likert-Skala online beurteilt.

Die dergestalt relativ differenziert abgefragten drei Evaluationsdimensionen lassen sich – stark abstrahiert – auch verstehen als prozedurale Vergewisserungskriterien für gute Guidelines zur Annotation literaturwissenschaftlicher (oder allgemein: geisteswissenschaftlicher) Konzepte auf Texten unterschiedlicher Genres. Sie können prozessorientiert abgebildet werden:

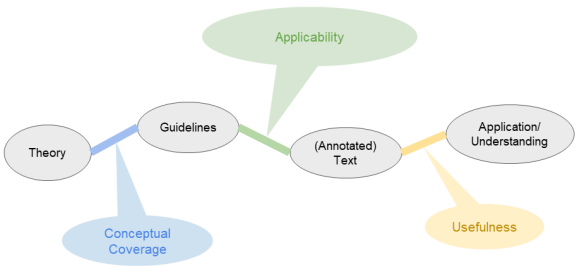

Abb. 1: Prozedurale Darstellung der Evaluationsdimensionen

Evaluation der Evaluation: Ergebnisse des Workshops

Am **dritten Workshop-Tag** wurden die Ergebnisse der Evaluation vorgestellt und diskutiert. Als methodisch ausgesprochen interessantes Resultat zeigte sich, dass die qualitative Plenumsdiskussion der Guidelines zu Einschätzungen führte, die durch die Resultate der quantitativen Evaluation in den Fragebögen abgebildet wurden: Die als theoretisch hochdifferenziert gelobten Guidelines waren just diejenigen, die in der ersten Dimension die meisten Punkte erzielten usw. Eine vierstufige Skala scheint also den gesamten Evaluationsbereich mit den drei komplexen Theorie-, Anwendungs- und Brauchbarkeitsfragen ausreichend differenziert abbilden zu können. Problematisch erschien den Teilnehmer/innen allerdings, dass die Fragen zunehmend schwerer zu beantworten waren. Dies resultierte aus der Schwierigkeit, für einige Fragen potentielle Anwendungsfälle zu antizipieren, in denen bereits annotierte Texte sinnvolle Forschungsfragen ermöglichen. Allerdings wurden im Gegensatz zur subjektiven Wahrnehmung der Teilnehmer/innen diese Fragen der letzten Dimension mit einer zunehmend geringeren Standardabweichung beantwortet. Trotz *gefühlt* größerer Schwierigkeiten mit den Fragen zur Usefulness wurden die Guidelines dort einvernehmlicher evaluiert.

Die drei Dimensionen erwiesen sich damit als praktikables Instrument einer differenzierten Bewertung der Guidelines. Das Experiment "Shared Task für die DH" ist also geglückt. Die drei Bewertungsdimensionen stehen allerdings auch weiterhin für eine der großen methodischen Herausforderungen im *Digital Humanities*-Bereich: die Evaluation von Operationalisierung, Analyse und Interpretation in interdisziplinären Kontexten.

Fußnoten

1. Siehe Gius et al. (2018), aber auch Reiter et al. (2017), bzw. zur Projekt-Dokumentation Gius et al. (2016ff.).
2. Bspw. Daniel Zeman et al. 2017 dokumentieren, wie ein typischer NLP Task funktioniert.
3. Wir danken der VolkswagenStiftung für die Finanzierung dieses Workshops, der vom 17.-19. Sept. 2018 an der Universität Hamburg stattgefunden hat. CRETA (Stuttgart) danken wir für die Finanzierung der notwendigen Annotationsarbeiten durch unsere HiWis Linda Kessler, Tanja Preuß, Nina Stark, Hanna Winter. Katharina Krüger hat uns bei der Organisation in Hamburg unterstützt und Carla Sökefeld den Workshop protokolliert.
4. Ausführlichere Informationen und Literaturhinweise zu einführender (etwa Jahn 2017), grundlegender (etwa Ryan 1991 und Genette 1980: 227-237) und vertiefender (etwa Mani 2013) narratologischer Literatur finden sich auf der Projekthomepage https://sharedtasksinthedh.github.io/levels
5. Die ausführliche Dokumentation der Abläufe des STs, die Publikation der Guidelines inkl. Reviews wird in zwei Sonderheften der *Cultural Analytics* publiziert (vsl. Q4/2018 und Q3/2019).
6. Alle Guidelines wurden 1) durch die jeweiligen Autor/innen selbst, 2) durch ein zufällig ausgewähltes anderes teilnehmendes Team und 3) durch von uns eingesetzte Hilfskräfte annotiert. Grundlage der Annotation waren acht literarische Prosatexte unterschiedlichen Umfangs, die zwischen 1797 und 1931 publiziert wurden. Als Metrik für das Agreement wurde Gamma verwendet (Mathet et al., 2015).

Bibliographie

Genette, Gérard (1972): *Narrative Discourse. An Essay in Method*. Ithaca 1980. (Franz. Figures III. Paris 1972).

Gius, Evelyn, / Nils Reiter / Marcus Willand (2016ff.): *Shared Tasks in the Digital Humanities. Systematic Analysis of Narrative Texts through Annotation*, Projektwebsite und -dokumentation https://sharedtasksinthedh.github.io/ [letzter Zugriff 28 September 2018].

Gius, Evelyn, / Reiter, Nils / Strötgen, Jannik / Willand, Marcus (2018): *SANTA: Systematische Analyse Narrativer Texte durch Annotation*. DHd2018, Köln.

Mani, Inderjeet (2013): *Computational Narratology*. Peter Hühn, John Pier, Wolf Schmid und Jörg Schönert (Hrsg.). The living handbook of narratology. Hamburg University Press http://www.lhn.uni-hamburg.de/article/computational-narratology [letzter Zugriff 28 September 2018].

Pier, John (2014): *Narrative Levels* (revised version; uploaded 23 April 2014). Peter Hühn, John Pier, Wolf Schmid und Jörg Schönert (Hrsg.). The living handbook of narratology. Hamburg University Press ttp://www.lhn.uni-hamburg.de/article/narrative-levels-revised-version-uploaded-23-april-2014 [letzter Zugriff 28 September 2018].

Reiter, Nils / Gius, Evelyn, / Strötgen, Jannik / Willand, Marcus (2017): *A Shared Task for a Shared Goal - Systematic Annotation of Literary Texts*. DH2017, Montreal.

Ryan, Marie-Laure (1991): *Possible Worlds, Artificial Intelligence, and Narrative Theory*. Bloomington: Indiana University Press.

Mathet, Yann / Widlöcher, Antoine / Métivier, Jean-Philippe (2015): *The unified and holistic method gamma (#) for inter-annotator agreement measure and alignment*. Computational Linguistics, 41(3):437–479.

Zeman, Daniel et al. (2017): *Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. CoNLL http://www.aclweb.org/anthology/K17-3001 [letzter Zugriff 28 September 2018].