Granularity Perspectives in Modeling Humanities Concepts

Piroska Lendvai
Research Institute for Linguistics, Budapest, Hungary
Tilburg centre for Creative Computing, Tilburg University, Netherlands
piroska@nytud.hu

ABSTRACT
Our pilot study focuses on the issue of concept granularity, at the intersection of the Digital Humanities and Language Technology disciplines. Based on our own work and on contributions by several authors in the current volume, we would like to initiate the charting of approaches taken from different fields of computational linguistics to the analysis of higher-level content units, typically in folk tale texts, for establishing such units and their composition.

1. INTRODUCTION
The target of several ongoing national and international projects is to establish or agglomerate digital research infrastructures for communities that have so far been less exposed to computational tools for data processing and analysis. Humanities and Social sciences (HSS) disciplines, where primary sources for study are typically text-based (such as literature, law, history, philosophy, religion, ethnography, etc.) are one of these new areas whose traditional research activities are getting enhanced by new and faster methods of text analysis or synthesis. Directly applying already existing tools and frameworks from Human Language Technology (HLT), as well as developing their own, the Humanities community is increasingly engaged in emerging paradigms of the Digital Humanities (DH) discipline. Issues in this field are growing complex, however, as the DH community necessarily mingles with that of HLT.

DH is clearly more than ways and means to enhance the individual researcher’s ability to organize and navigate large amounts of data, but is rather a dynamic new field with a focus on exploring possibilities of linguistic and textual analysis, such as semi-automatic semantic annotation, thesauri, concordances (cf. [1]). Most importantly, applying and developing computational approaches for HSS data needs to take into account the legacy of needs and specificities of the types of research questions and research materials in these fields; the survey in [2] aims to provide an overview of these.

It appears that the complexity of typical HSS research issues, together with the modules of possibly applicable HLT solutions might be viewed as a matrix of layers, components, and granularity, in which it is not straightforward to identify correspondences in terms of which HSS and HLT research categories can or could be coupled – the latter being one of the core aims of DH. In recent initiatives, such as by the efforts of the CLARIN project1, linguistic analysis and processing tools, corpora and repositories holding linguistic information, as well as analysis and synthesis methods of HLT, all lying at the basis of such facilities, are being brought to HSS specialists. It has been noted that HSS scholars (just like all other non-HLT experts) are often unaware that automatic linguistic analysis is a prerequisite of higher-level content processing.

Based on the materials presented at the First International Workshop of the AMICUS project2 that focuses on automated motif discovery in cultural heritage and scientific communication texts, we aim to report on the approaches and viewpoints of HLT specialists with experience in DH, on the issue of compositionality appearing in their suggested approaches and models of Humanities concepts. The role of the authors of these studies, being creators of DH resources – as opposed to users and curators of these – is to plot the mechanisms that produce structures that can be understood as ‘motifs’, to map them to computational procedures, resources, or tools, connecting the HSS and HLT disciplines.

Such a pioneering venture has several challenging aspects, out of which we propose to pay attention to but a few ones:

- A certain concept or notion for one scientific field might be fundamental, and thus regarded as atomic, whereas in another field (serving a different purpose) it might or should be modeled as compositional;
- Cases when it is not yet possible or desirable to make some notion specific, i.e. explicit, resulting in underspecification;
- The mingling of pragmatic and semantic features, and modeling possibilities for interfacing these two traditional linguistic levels.

The goals of the authors of papers in this paper are different (analysis, generation, automated markup), but common underlying aims seem to be arriving at “psychologically compelling results” [Malec, this volume], or detecting as much structure as possible from raw texts in order to achieve content analysis (e.g. [Declerck and Scheidel, this volume]) or synthesis

---

1 http://www.clarin.eu
2 http://amicus.uvt.nl
(e.g. [Gervás, this volume]), utilizing semantic concepts from both the HSS and the HLT communities.

2. ADDRESSING FOLK TALE TEXTS IN TERMS OF HLT

The goal of the AMICUS project is to distill knowledge about constructs from texts that seem weakly structured to a machine or to non-experts in (Digital) Humanities. It is possible for skilled humans to infer semantic, or even subtle cognitive structures in these, nevertheless, such mechanisms are even manually difficult to capture by one of the mainstream methodological approaches of natural language processing (NLP), for example, in terms of identifying conceptual boundaries in text (e.g. finding names of heroes in tales, or Proppian functions), assigning category labels to these, and arranging them in a (hierarchical) structure. Once such tasks are solved, automatic annotation is easier to design and enhance by means of picking cues for determining (sequence) similarity, and feeding in structured resources.

To address this gap between HLT and human cognitive processes that produce folklore texts or argumentative presentation of research results, the metaphor AMICUS operates with is that of a motif. The project faces the challenge of investigating whether a motif is an underspecified concept or mechanism as such, and whether this property of motifs contributes to or enables devising better HLT models in the HSS fields, or not.

2.1 Semantic and Pragmatic Levels

It turns out from our investigations during the first half-year of the AMICUS project that working with motifs entails examining narration by which they exhibit themselves in text. Just like all types of material carried by spoken or written language, narrative on the higher linguistic levels operates by pragmatics and semantics. Features of the pragmatic level are expressed by properties of discourse, carrying communicative goals, reflected e.g. by stylistic markers and arrangement of the content. The semantic level pertains to content: the kind of entities, facts, and events appearing, but, importantly, also the abstract domain knowledge expressed as categories – such as the notions of ‘entity’, ‘fact’, and ‘event’.

It appears from, although not stated by, the systematic survey of [Gervás, this volume] that narrative operates by means of both low (i.e. easily detectable by machines) and high level features (i.e. difficult to formalize) on both the semantic and pragmatic levels. An important HLT issue with respect to this, but unclear at the moment, is if it is possible to classify these features and concepts as pertaining to – or, modellable by – exclusively the semantic or the pragmatic level. We note that this is inherent in language, applicable to possibly several, if not most, HLT levels (investigated by us in [3] in the HLT domain of dialogue processing). However, the issue now is whether treating this as a general strategy, i.e. (manifesting itself as) underspecification rather than compositionality, brings one closer to the goals of AMICUS.

The strategy of underspecification (or, opacity) here is that in fact (some) pragmatic and (some) semantic phenomena co-depend on one another. This might entail that separating those by grouping under different linguistic levels and labels is a convenience issue, by now a burdensome legacy, suboptimal in the DH discipline. In other terms, such modeling could pose an obstacle to optimal processing and might risk obtaining valid DH models from and creating valid approaches to Humanities data.

In the following section we will illustrate this line of thought by examples.

2.2 Matrix of correspondences

In Table 1 we attempt to initiate the sketch of a matrix, quoting concepts and the terms used to denote them as suggested in some of the studies published in the current Proceedings, as well as arranging them in a simple structure, reflecting a lightweight and non-systematic hierarchy. The goal is to align related concepts occurring in some of the research papers as belonging to the so-called pragmatic concepts (top section), respectively to the semantic concepts (bottom section). Sometimes these denote similar phenomena or perspectives, expressed by different terms, sometimes they complement one another.

With respect to compositionality, it is clear that the picture is extremely complex, and, judged already by the fact that examples for pragmatic features are only possible to express by means of involving/quoting actual content elements, it seems to be indeed the case that pragmatic and semantic phenomena co-occur in intertwined ways. Events and event sequences are high-level content conglomerates, as we noted above, since the typical elements involved (time, location, action, means, etc.) are themselves possibly composed of multiple (linguistic) entities; the same goes for motifs. For example, a person in history or a character’s name in a fairy tale in fact refers to a single entity, but in order to be able to represent e.g. ‘Churchill’, we need to formalize a range of properties such as first name, gender, etc. in order to disambiguate, deduplicate references, link data, and so on. The same mechanism applies to many other concepts, such as time – but not necessarily in all genres: fairy tales do not involve explicit dates.

In Table 1 we additionally attempt to signal the difficulty of the required linguistic processing level, which would indicate the feasibility of possibly applicable HLT methodologies: manual or automated. It is common practice that content-related phenomena can be successfully dealt with by gazetteer approaches (i.e., identifying the entities based on a fixed list), such as in Information Extraction (IE). It is to be judged on the basis of actual implemented procedures whether such semantic concepts would/could be subsequently returned to for further specification and/or revision, once the pragmatics involved in expressing them has been identified and separated from content, keeping them underspecified for a while (cf. the approach sketched in [Declerck, this volume]).

In general, since only part of the phenomena listed in Table 1 have been formalized so far, we would like to invite all interested parties to contribute to filling in this, or a similar, matrix, thereby benefitting from input given by representatives of all communities related to analysis of high-level conglomerates in narrative genres. It requires thorough investigation if the sketched matrix would apply to both folk tales (and other literary genres) as well as to scientific communication texts. The table of correspondences in [de Waard, this volume], comparing story grammar elements with syntagmatic components of scientific texts, contains (high-level) data categories and ideas that are readily importable.
The matrix could suggest intersections identified between HLT and DH-HSS, and one may understand in more detail which aspect of narration and motif are carried by which formalizable linguistic levels. We also propose to prepare a gold-standard example set marked up by the relevant phenomena.

3. MOTIFS

Motifs are an example of the vehicles of the semantics of communication, for example in folk tales, but possibly in all kinds of discourse. They contain sets of content objects in a predefined way, the segmentation and granularity of which are perhaps roughly equally distributed within a story. It seems that even when some heavy artillery of HLT is utilized, motifs remain troublesome to address, since establishing and computing the segment boundaries of their constituents proves to be complex. The work initiated within the AMICUS network, receiving its initial form in [4], will hopefully alleviate some of the problems related to this. In this section we propose to tackle some further aspects, possibly less addressed in previous research of motifs.

It would be interesting to examine which constituent of a motif should change in order to require the update of the motif – which is in fact the moving on to a new motif. Establishing such a criterion might be an indicator for motif segmentation in certain constellations.

Table 1. Pragmatic features (top section) and semantic concepts (bottom section) in approaches to (structural) narrative analysis (cf. Propp) and generation (cf. Gervás), information extraction (cf. Declerck), machine learning (cf. Malec), access to knowledge from scientific papers (cf. de Waard). We indicate the typical linguistic processing level that could be associated with each object in a linguistic processing module.

<table>
<thead>
<tr>
<th>Analysis feature</th>
<th>Linguistic manifestation e.g.</th>
<th>Example and possible cue</th>
<th>Linguistic analysis level (difficulty: Low, Medium, High)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mood</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voice: Narrator present/absent (IE: “speaker”) Narrator’s Function (neutral, directing, ideological…)</td>
<td>Cue words, Grammatical person</td>
<td>“Dear children”</td>
<td>Lexical (d:L)</td>
</tr>
<tr>
<td>Narrate order (Prior, simultaneous, subsequent, interpolated) Presentational ordering</td>
<td>Temporal expressions Timeline (underspecified)</td>
<td>Event has happened before/is going to happen after narration</td>
<td>Pragmatic/Semantic interface (d:M/H)</td>
</tr>
<tr>
<td>Speed/Pace</td>
<td>Event’s descriptive granularity</td>
<td>Summary of what happened vs (quoting) a dialogue</td>
<td>Semantic/Cognitive (d:M/H)</td>
</tr>
<tr>
<td>Frequency of mentioning an event</td>
<td>Repetition, Paraphrase, Summary</td>
<td>…</td>
<td>(d:L/M)</td>
</tr>
<tr>
<td>Perspective</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Focalization: Zero, internal, external Focalization on individual characters</td>
<td>Narrator knows more/equal/less than characters First person singular</td>
<td>…</td>
<td>Cognitive (d:H)</td>
</tr>
<tr>
<td>Program</td>
<td>textual unit that moves the story forward, irrespective of the dramatis personae</td>
<td>Negation, imperative</td>
<td>“Do not leave the house”</td>
</tr>
<tr>
<td>Event</td>
<td>abstract schema of systematic constituent structure</td>
<td>Cue words for Goal, Subgoal, Method…</td>
<td>“the aim of this study is…”</td>
</tr>
<tr>
<td>Entry</td>
<td>IE: - Character - Sets of Characters of same type?</td>
<td>Gazetteer, coreference, family relations from ontology</td>
<td>Attributes: age, synonyms?</td>
</tr>
<tr>
<td>- Object</td>
<td>Gazetteer, cue phrases</td>
<td>apple</td>
<td>Lexical (d:L)</td>
</tr>
<tr>
<td>- Sets of objects of same type</td>
<td>Gazetteer, cue phrases</td>
<td>apple</td>
<td>Lexical (d:L)</td>
</tr>
<tr>
<td>A. is</td>
<td>Profile/adjective</td>
<td>gazetteer</td>
<td>golden</td>
</tr>
<tr>
<td>event</td>
<td>When, Where, WhatAction, who to Whom, Why, How</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>
We would also like to understand in what ways motifs (not necessarily in the Proppian sense) pertain to the pragmatic level. Clearly, the ‘meaning’ of a motif is more than the agglomerate of its elements. By a similar mechanism, the units ‘youngest’ + ‘son’ would equal to ‘Character: hero’. It would be necessary to put into context the underlying formalism of cognitive processes in such cases.

4. INTERDISCIPLINARITY

In DH studies typically some close, goal-oriented research collaboration takes place between the language technology experts and the HSS researchers, where distinct disciplines come together, for example philosophy or sociology, as well as computational linguistics. It is important to ensure that the knowledge of the different communities involved is in such cases as optimally merged as possible in order to facilitate research and to maximize the success of the given project. One crucial aspect in this is that specialists belonging to the different communities need to make sure they understand each other's goals, methods, and, last but not least, terminology. This requirement however necessitates obtaining several insights about a new domain for all parties, which can be realized only by consciously investing time and effort in this process.

There is a clear need at several phases of a research project that representatives of different communities engage in in-depth dialogue with each other in order to gain new knowledge about each other's research foci, priorities, domain concepts, and language use. The establishment and practical implementation of this important layer of collaboration call for the identification of prerequisites that allow for optimally managing and realizing transdisciplinary research, involving simultaneous coordination and evaluation of steps to be taken during what we see as recursive exposure to knowledge from a new domain for both HSS and HLT specialists.

In the current section we would like to raise some issues, which, when answered, could possibly take us closer to our goal of integrating HLT into HSS. For example, it is an intriguing question whether some HLT methods can be readily utilized or adapted for fast modeling of a HSS domain, i.e. for knowledge acquisition and representation, in order for the HLT specialist to obtain a quick overview of it, as an alternative for reading up on a domain from textbooks and publications? Possibly, IE is capable of supplying such a method, semantic technologies likewise.

We also suppose that HLT experts need to be aware of the nature of typical HSS research methodologies. For example, it is not known if the presence of tools for automatically processing data on a certain linguistic level implies a must for applying the tool, even if HSS researchers initially think this information is irrelevant for their project.

We expect such questions to be answered as new research paradigms emerge from cross-fertilization among communities.

5. ACKNOWLEDGMENTS

Our thanks to Thierry Declerck for comments on the draft of this paper.

6. REFERENCES


