Which Dimensions of Narratives are Relevant for Human Judgments of Story Equivalence?

Bernhard Fisseni[1,2,3], **Benedikt Löwe**[1,3,4,5]

[1] Institute for Logic, Language and Computation, Universiteit van Amsterdam
Postbus 94242, 1090 GE Amsterdam, The Netherlands
b.loewe@uva.nl
[2] Fakultät für Geisteswissenschaften, Universität Duisburg-Essen
Universitätsstraße 12, 45117 Essen, Germany
bernhard.fisseni@uni-due.de
[3] Isaac Newton Institute for Mathematical Sciences
20 Clarkson Road, Cambridge CB3 0EH, United Kingdom
[4] Fachbereich Mathematik, Universität Hamburg
Bundesstraße 55, 20146 Hamburg, Germany
[5] Corpus Christi College, University of Cambridge
Cambridge CB2 1RH, United Kingdom

Abstract

We present an experimental approach to determining natural dimensions of story comparison. The results show that untrained test subjects generally do not privilege structural information. When asked to justify sameness ratings, they may refer to content, but when asked to state differences, they mostly refer to style, concrete events, details and motifs. We conclude that adequate formal models of narratives must represent such non-structural data.

1. Introduction

Traditional and current computational models of narrative (Rumelhart, 1980; Lehnert, 1981; Schank, 1982; Dyer, 1983; Turner, 1994; Pérez y Pérez and Sharples, 2004; Frank et al., 2003; Mateas and Stern, 2003; Mueller, 2004; Si et al., 2005) focus on structural aspects of the narrative in their representation: events, causal relations between events, temporal relations of events, agents of the narrative, spatial relations between agents and objects, etc.

In terms of the narratological distinction of *story* and *discourse* (cf., e.g., (Chatman, 1980)), the formal representation of the narrative is stressing the *story* over *discourse* (cf. (Young, 2007)). Several approaches strongly emphasize that abstracting from the *discourse* results will yield the *structural core of the narrative* that is used for storing the narrative in their memory, retelling the narrative, as well as decisions whether two narratives are the same. As a motivation for her *Plot Units*, Lehnert connects them to the cognitive representation of summaries in the mind of the reader: "When a person reads a narrative story, an internal representation of that story is constructed in memory." (Lehnert, 1981, pp. 293)

The most prominent examples of this approach are *Structure Mapping Theory* and its implemented version, the *Structure Mapping Engine* (Gentner, 1983; Falkenhainer et al., 1989). While technically not about narratives, but about analogy and analogical reasoning, its most iconic examples are the comparisons of narratives like *Karla the Hawk* and modifications (Gentner et al., 1993, p. 533) and the question whether human test subjects recognize structural analogy:

> Domains and situations are psychologically viewed as systems of objects, object-attributes, and relations between objects. ... These representations ... are intended to reflect the way people construe a situation. (Gentner, 1983, p. 156–157)

This emphasis on structural analogy is reflected in recent approaches to find formal representation systems for comparison of narratives (Löwe, 2010; Bod et al., 2011; Löwe, 2011; Bod et al., 2012). On the other hand, systems based on *Structure Mapping Theory* have been criticised on the basis of empirical results for ignoring salient features of narratives that are relevant for human judgments of story equivalence (cf. also the discussion in (Löwe, 2011, § 2)):

> We have shown that [the] lack of inclusion of emotive content [in Gentner's *Structure Mapping Engine*] has made it psychologically implausible. (Lam, 2008, p. 38)

A natural and much more general follow-up question is addressed in this paper, namely: *Which features of narratives are relevant for human judges of story equivalence?*

In other words: do untrained human subjects, confronted with the task of deciding whether narratives are "the same" (without further specification what is the precise meaning of this phrase), rely mostly on structural features, or do other features (that are traditionally counted as part of the *discourse*) play a role in these decisions as well?

Since we do not want to presuppose any particular narratological ontology of features and their classification, we use the vague term *dimension* to refer to the possible features of narratives that could potentially be used to distinguish narratives as similar or equivalent. Examples of potential *dimensions* are: (a) motifs and superficial aspects such as (features of) the setting, the inventory of characters, single

events (not connections between events) stylistic similarity, but (b) also aspects story event structure. Other, more philological categories would be (c) the relationship between narrator, characters, reality, the "possible world" of the story (cf., e.g., (Martinez and Scheffel, 2009)). We consider it part of the goal of the research reported on in this paper to give a preliminary classification of the relevant *dimensions* as they occur in the empirical data.

In § 2., we give a description of three experiments that elicited story comparisons; in § 3., we discuss the results of the experiments and conclude with a list of ideas for future work.

2. Experimental Work

The experimental work approached the question how test subjects naturally talk about stories when comparing them. Hence, neither the structural nor the motif level were focused by the instructions. As stimuli, we used stories that were not specifically constructed for the purpose of the experiment, but at most slightly varied to introduce controlled differences.

The tasks were simple and the experiments were conducted as classroom experiments. Participants were students of German literature and language and were rewarded for their participation with chocolate. Test subjects received a questionnaire with instructions; these were also presented by the experimenter. Test subjects were given about 15 minutes (**Queneau I** and **Queneau II**) or 20 minutes (**Fairy Tales**) to perform the task. Only numbers for native speakers are reported, unless stated otherwise. In all experiments, test subjects were given ample opportunity to report difficulties and give commentaries.

2.1. Experiment *Queneau I*

Setup. Test subjects were given two of Queneau's *Exercises in Style* (Queneau, 1947), translated into German (Queneau, 1990) with a length between 7 and 12 lines. Queneau's work consists of 99 variants of a base story in which the narrator gets on a bus, witnesses an altercation between a man and another passenger, and then sees the same person later getting advice on adding a button to his overcoat. We selected the base variant (*notations*), the variant told in reverse temporal order (*rétrograde*), the variant in which the agents are replaced by botanical objects (*botanique*), and the variant in which offensive language is used (*injurieux*). In the following, we refer to the variants as **order**, **botan.** and **offens.**, respectively.

Each test subject was given the base variant and one of the other variants; the title of the stories and their provenance was not given.[1] Test subjects were asked to take the role of the editor of a story magazine, helping a colleague to make a decision with respect to a strict rule of the journal: not to publish the same story twice. The test subjects should determine whether the two stories given were the same, and should explain to their colleague why they reached this conclusion. It was varied whether the colleague himself had suggested that the stories were the same or not.[2]

[1] One test subject recognised the stories.

[2] This turned out to have no observable effect on the test subjects' judgment.

There were 65 test subjects overall, of these 59 native speakers of German, almost all in their first semester and most of them studying to become primary school teachers.

The responses of the test subjects were categorised *ex post*, and the frequency of the categories was reviewed; from the natural language descriptions used by the test subjects, 46 *labels* were constructed (most occurring very infrequently) and later grouped into eight *categories*: *content*, *details*, *imagery*, *order*, *structure*, *style*, *substitution*, and *theme*. For every story, there were categories that was expected to figure most prominently: *order* for variant **order**, *substitution* or *imagery* for variant **botan.**, and *style* for variant **offens.**

	order	**botan.**	**offens.**
Same story	17	8	10
Different stories	3	8	11
(no decision)	0	2	0
n	20	18	21

	order	**botan.**	**offens.**
Same story	4	4	3
Different stories	2	1	6
mentioned	6	5	9

Table 1: Sameness judgments and *expected categories* by test subjects (**Queneau I**). The upper table lists the judgments as *the same story* and *different stories* by variants. The lower table lists how many times the *expected categories* are mentioned as a factor of difference.

	details			**structure**		
sameness	**simil.**	**diff.**	*no dec.*	**simil.**	**diff.**	*no dec.*
order	2	1	0	0	0	0
botan.	1	3	1	2	1	0
offens.	1	2	0	0	0	0
	content			**theme**		
sameness	**simil.**	**diff.**	*no dec.*	**simil.**	**diff.**	*no dec.*
order	11	3	0	2	1	0
botan.	5	5	0	1	1	0
offens.	9	7	0	4	2	15

Table 2: Mention of *structure* and *details*, *content* and *theme* per story (**Queneau I**)

Results. In all cases, the *expected categories* are mentioned by a minority of test subjects (cf. Table 1): For the variant **order**, only six test subjects mention *order* as a factor of difference; for the variant **botan.**, *substitution* is mentioned by two people, *imagery* by five; for the variant **offens.**, *style* is mentioned by nine test subjects. This is particularly striking in the case of a difference of order, where the vast majority of test subjects considers the stories to be *the same*.

The categories *details* (place, time, location, etc.), *structure* (surface structure, deep structure, etc.), *theme* and *content* occur very rarely (cf. Table 2). *Content* and *theme* are used as an argument in favour of similarity in nearly all cases when they are mentioned (with two exceptions for *theme*). Only three people mention *structure* at all.

Difficulties. Test subjects do not generally formulate their answers clearly and assigning the categories to the descriptions requires interpretation of the intention of the test sub-

	order	botanic	offensive
same story	8	0	3
different story	0	2	5
no decision	0	0	7
n	20	18	21

Table 3: Test subjects mentioning *perspective* as a differing factor by decision regarding story similarity (**Queneau I**).

jects. As an example, we mention the use of the label "perspective" illustrated in Table 3 (only noted as a factor of difference). The numbers suggest that the test subjects have a very vague notion of perspective.

Interpretation. We interpret these data to show that structural factors are not the most important aspect with respect to which test subjects compare stories, if this is not triggered explicitly. We are surprised that the *expected categories* are not named more often.

2.2. Experiment *Queneau II*

The main change from the previous experiment was that we intended to increase the number of mentioned categories of comparison per test subject, with the expectation that this will increase the number of mentions of structural factors and the *expected categories*. Test subjects were asked to justify their decision regarding the sameness of the story by naming at least two "important aspects" with respect to which the stories were the same or differed. As in **Queneau I**, test subjects had to take the role of the editor; the additional layer of communicating to a colleague was removed. Stories and questionnaires remained the same, only that the order of different and same aspects was varied, without any effect. 41 test subjects, 37 of them native speakers, participated; most were in their second year and intending to become a teacher at a grammar or comprehensive school.

Results. Explicitly asking for more than one category had a strong effect: Nearly all subjects (30 out of 37) now mention single instances of events or details[3] regarding settings and characters in their lists (21), or textual details like text length (13), both as similarities or differences (often different categories for either side). The category *content* is again mentioned by about half the subjects (cf. Table 5), always as a similarity, and *theme* is again rather rare (cf. Table 5), but except for two cases in variant **offens.** it is mentioned as a difference. The *expected categories* are now mentioned by a majority of test subjects for each variant (cf. Table 4).
Only two test subjects formulate their observations regarding *order* identifying the temporal order of the second story;[4] the other four formulate in a way that it is not clear whether they correctly resolved the order of events, or assumed a reverse chronological order. Regarding variant **botan.**, most test subjects present the observation in very concrete terms ("humans and vegetables") rather than abstractly.

	order	botan.	offens.
Same story	6	2	7
Different stories	5	1	6
(no decision)	0	9	1
n	11	12	14
factor of...	**order**	**botan.**	**offens.**
Same story	0	0	0
Different story	6	8	10
mentioned	6	8	10

Table 4: Sameness judgments and *expected categories* by test subjects (**Queneau II**). The upper table lists the judgments as *the same story* and *different stories* by variants. The lower table lists how many times the *expected categories* are mentioned as a factor of similarity or difference.

	details			textual details		
sameness	**simil.**	**diff.**	*no dec.*	**simil.**	**diff.**	*no dec.*
order	5	2	0	1	2	0
botan.	1	5	1	1	4	0
offens.	3	3	1	4	2	0
	content			**theme**		
sameness	**simil.**	**diff.**	*no dec.*	**simil.**	**diff.**	*no dec.*
order	7	0	0	0	0	0
botan.	0	2	1	3	1	0
offens.	6	0	0	2	1	0

Table 5: Mention of *details*, *textual details*, *content* and *theme* per story (**Queneau II**)

The vast majority of "important differences" reported (several for nearly all test subjects) were details and motifs such as places and characters. A preference for reporting structural similarities could again not be confirmed.[5]

2.3. Experiment *Fairy Tales*

The preceding experiments used variants of a very short story with a very limited structure. As a next step, we aimed at testing story comparison with stories of greater structural complexity.

Stimuli. Each test subject was given two versions of the fairy tale *Die drei Federn* (*The three feathers*) of the Brothers Grimm. The base version was the short version from the first edition (Grimm and Grimm, 1812, No. 64, III) and the variants were versions of the significantly altered and longer version from the last edition (Grimm and Grimm, 1857, No. 63) altered in several ways (see below). In the story, a king sets tasks for his sons to complete; the *Dummling* (Stupid One) completes all of them, with magical help, while his brothers fail. This story was used because the two versions

[3] For reasons of space, we cannot give a complete breakdown of the data, but give only one example of a questionnaire (variant: **order**, decision: **same story**). The test subject mentions as aspects of similarity: *"Ort: Autobus, Gare Saint-Lazare; Zeit: Mittag; Personen: Mann mit Hut, Freund von jenem; Detail: Hut, Handlung, Mann im Bus, mit Hut"*. ("Location: bus, Gare Saint-Lazare; Time: Noon; Characters: Man with hat, friend of his; detail: hat, plot, man on the bus, with hat"); he or she mentions as aspects of difference: "[In] Gesch. 2 ist Zeit („heute Mittag") genau erwähnt" ("In story 2, the time is mentioned precisely ('this afternoon')").

[4] We counted a mention of "temporal perspective" as an identification of the reversed order of narration.

[5] The following is an interesting quotation from one of the test subjects' answers: „Der Inhalt macht keine Geschichte aus; es kommt auf die Darstellungsweise und die benutzten Mittel an." ("Content does not determine a story; it is about the way of presentation and the [stylistic] devices used.")

were (to the experimenters) immediately recognisable as the same story, but were quite different in many ways: from details of the story to the concrete kind of the tasks.

There were four variants of the second version of the story:

Temp: the original version with a slight variation of the order to presentation: the outcome of the tasks was recounted before the details of the tasks.

Granularity: a version with lower granularity (similar to a summary),

End*: versions with a different ending, namely **End1:** one in which the end was just reversed (the brothers ruled, haggling until their death), **End2:** one in which the *Dummling* rules, but is remembered for bad governance and stupidity.

Setup. The instructions asked test subjects to report at least three "important aspects" with respect to which the stories were similar or differed. It was expected that test subjects would report a variety of differences, both structural and details. Test subjects were 38 students of German literature and language, most of them in their first year.

Results and interpretation. More than in **Queneau II**, test subjects overwhelmingly name details. The categories *content* or *structure* are very rare (six occurrences overall, three non-native speakers, mentioned as "similar" in all cases). There are also just two mentions of the category *style*.

	Temp	**Granularity**	**End1**	**End2**
Same story	12	7	3	2
Different stories	2	5	4	2
(*no decision*)	0	0	1	0
story is ...	**Temp**	**Granularity**	**End1**	**End2**
Same story	3	1?	2	1*
Different stories	1	0	1	2
(*no decision*)	0	0	1	0

Table 6: Sameness judgments and mentions of the *expected labels* by test subjects (**Fairy Tales**). The upper table lists the judgments as *the same story* and *different stories* by variants. The lower table lists how many times labels corresponding to the actual manipulations are mentioned as a factor of difference. In the lower table, "?" marks an uncertain classification; "*" indicates that the end is mentioned, but the test subject wrongly claims that the ending is the same in both variants.

In the lower part of Table 6, we list how many test subjects recognized the actual manipulations of the stories. The data show that our manipulations do not generally result in the judgment that the stories are different. Differences between the stories are – according to the extension of Fisher's exact test for data as implemented in R (R Development Core Team, 2010) – at best marginally significant.

The vast majority of "important differences" reported (several for nearly all test subjects, while only few test subjects mention structural factors such as "course of action") were details and motifs such as places and characters. Relatively few test subjects mention the factors we manipulated in the story (cf. Table 6, lower half).

Seven people claim to know at least one of the stories; one of them claims to know both, clarifying: "→ same tale"; another modifies: "more or less", another: "parts of it, *Froschkönig*, *Aschenputtel*" (the Frog Prince, Cinderella), and also two of those who do not know the stories, say: "parts of it from other stories" or "the tale of *Aschenputtel*" (Cinderella). These remarks confirm that test subjects have a mixed motif-structure view on these tales.[6]

3. Discussion & Conclusion

We conclude that structural dimensions of stories are not a natural level of processing sameness judgments for untrained subjects. Different tasks trigger different reactions by test subjects: When asked to justify their actions, test subjects may refer to a vague notion of content, which arguably encompasses the event structure and causal links. However, when asked to produce many factors, references become much more concrete and less structural.[7] Details and motifs, linguistic features and other dimensions are also used by test subjects.[8]

We conclude that our experiments may be seen as evidence that either structural similarity does not suffice for sameness judgments or that the empirical grounding for formal models of narrative should not be based only on untrained and unfocused subjects. If a model of story similarity is to be cognitively adequate for untrained and unfocused subjects, it must allow selective access according to the goal of the comparison, and must be complemented by a model of story processing that determines which dimensions are focused.

Acknowledgements

The research in this paper was funded by the *John Templeton Foundation* (JTF) via the project *What makes stories similar?* (grant id 20565) and the *Nederlandse Organisatie voor Wetenschappelijk Onderzoek* (NWO) via the project *Dialogical Foundations of Semantics* in the ESF EuroCoRes programme LogICCC (LogICCC-FP004; DN 231-80-002; CN 2008/08314/GW). The authors acknowledge the financial support and the kind hospitality of the *Isaac Newton Institute for Mathematical Sciences* (programme *Semantics & Syntax*). The authors should like to thank Tim Kocher and Charlotte Wollermann (Duisburg-Essen) for giving them access to the test subjects.

4. References

Rens Bod, Benedikt Löwe, and Sanchit Saraf. 2011. How much do formal narrative annotations differ? A Proppian case study. In Charles Ess and Ruth Hagengruber, editors, *The computational turn: Past, presents, futures?, Aarhus University, July 4–6, 2011*, pages 242–245, Münster. MV-Wissenschaft.

[6] Consider also the following quotations from the questionnaires: „Die Geschichte hat den gleichen Kern, ist äußerlich aber unterschiedlich." ("The story has the same core, but is different in outer appearance"; judgment: different story, **Temp**) and „Die Intention bleibt auch bei leicht abgewandeltem Inhalt gleich (letzter Satz)." ("The intention stays the same with [slightly] altered content (last sentence)"; judgment: same story, **Temp**).

[7] It is also conceivable that a different educational background of the test subjects, could have a significant effect on the results.

[8] Motif and story indices (Uther, 2004; Thompson, 1955 1958) highlight that different levels of story comparison are philologically interesting.

Rens Bod, Bernhard Fisseni, Aadil Kurji, and Benedikt Löwe. 2012. Objectivity and reproducibility of Proppian narrative annotations. this volume.

Seymour B. Chatman. 1980. *Story and Discourse: Narrative Structure in Fiction and Film*. Cornell University Press.

Michael G. Dyer. 1983. *In-depth understanding: A computer model of integrated processing for narrative comprehension*. Artificial Intelligence Series. MIT Press, Cambridge MA.

Brian Falkenhainer, Kenneth Forbus, and Dedre Gentner. 1989. The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, 20:1–63.

Stefan L. Frank, Mathieu Koppen, Leo G. M. Noordman, and Wietske Vonk. 2003. Modeling knowledge-based inferences in story comprehension. *Cognitive Science*, 27:875–910.

Dedre Gentner, Mary Jo Rattermann, and Kenneth D. Forbus. 1993. The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology*, 25:524–575.

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170.

Jacob Grimm and Wilhelm Grimm. 1812. *Kinder- und Hausmärchen*. Realschulbuchhandlung, Berlin, 1st edition.

Jacob Grimm and Wilhelm Grimm. 1857. *Kinder- und Hausmärchen*. Dieterichsche Buchhandlung, Göttingen, 7th edition.

Samantha Lam. 2008. Affective analogical learning and reasoning. Master's thesis, School of Informatics, University of Edinburgh.

Wendy G. Lehnert. 1981. Plot units and narrative summarization. *Cognitive Science*, 4:293–331.

Benedikt Löwe. 2010. Comparing formal frameworks of narrative structures. In Mark Finlayson, editor, *Computational models of narrative. Papers from the 2010 AAAI Fall Symposium*, volume FS-10-04 of *AAAI Technical Reports*, pages 45–46.

Benedikt Löwe. 2011. Methodological remarks about comparing formal frameworks for narratives. In Patrick Allo and Giuseppe Primiero, editors, *Third Workshop in the Philosophy of Information, Contactforum van de Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten*, pages 10–28, Brussel. KVAB.

Matias Martinez and Michael Scheffel. 2009. *Einführung in die Erzähltheorie*. C. H. Beck, München, 8th edition.

Michael Mateas and Andrew Stern. 2003. Integrating plot, character and natural language processing in the interactive drama *Façade*. In Stefan Göbel, Norbert Braun, Ulrike Spierling, Johanna Dechau, and Holger Diener, editors, *Technologies for Interactive Digital Storytelling and Entertainment. TIDSE 03 Proceedings*, volume 9 of *Computer Graphik Edition*. Fraunhofer IRB Verlag.

Erik T. Mueller. 2004. Understanding script-based stories using commonsense reasoning. *Cognitive Systems Research*, 5(4):307–340.

Rafael Pérez y Pérez and Mike Sharples. 2004. Three computer-based models of storytelling: BRUTUS, MINSTREL and MEXICA. *Knowledge-Based Systems*, 17:15–29.

Raymond Queneau. 1947. *Exercices de style*. Gallimard, Paris.

Raymond Queneau. 1990. *Stil-Übungen*. Suhrkamp, Frankfurt a.M.

R Development Core Team, 2010. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Wien.

David E. Rumelhart. 1980. On evaluating story grammars. *Cognitive Science*, 4:313–316.

Roger C. Schank. 1982. *Dynamic memory: A theory of reminding and learning in computers and people*. Cambridge University Press.

Mei Si, Stacy C. Marsella, and David V. Pynadath. 2005. Thespian: Using multi-agent fitting to craft interactive drama. In Michal Pechoucek, Donald Steiner, and Simon Thompson, editors, *AAMAS '05. Fourth International Joint Conference on Autonomous Agents and Multiagent Systems 2005. Utrecht, Netherlands. July 25–29, 2005*, pages 21–28. ACM.

Stith Thompson. 1955–1958. *Motif-index of Folk-Literature: a Classification of Narrative Elements in Folktales, Ballads, Myths, Fables, Medieval Romances, Exempla, Fabliaux, Jest-Books, and Local Legends. 6 volumes.* Indiana University Press, Bloomington, 2nd edition.

Scott Turner. 1994. *The creative process. A computer model of storytelling*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Hans-Jörg Uther. 2004. *The Types of International Folktales. A Classification and Bibliography. Based on the System of Antti Aarne and Stith Thompson*. The Finnish Academy of Science and Letters, Helsinki.

R. Michael Young. 2007. Story and discourse: A bipartite model of narrative generation in virtual worlds. *Interaction Studies*, 8:177–208.