Narrative Similarity as Common Summary

Elektra Kypridemou and Loizos Michael

School of Pure and Applied Sciences
Open University of Cyprus
Nicosia, Cyprus
elektra.kypridemou@st.ouc.ac.cy, loizos@ouc.ac.cy

Abstract

The ability to identify similarities between narratives has been argued to be central in human interactions. Previous work that sought to formalize this task has hypothesized that narrative similarity can be equated to the existence of a common summary between the narratives involved. We offer tangible psychological evidence in support of this hypothesis. Human participants in our empirical study were presented with triples of stories, and were asked to rate: (i) the degree of similarity between story A and story B; (ii) the appropriateness of story C as a summary of story A; (iii) the appropriateness of story C as a summary of story B. The story triples were selected systematically to span the space of their possible interrelations. Empirical evidence gathered from this study overwhelmingly supports the position that the higher the latter two ratings are, the higher the first rating also is. Thus, while this work does not purport to formally define either of the two tasks involved, it does argue that one can be meaningfully reduced to the other.

1998 ACM Subject Classification J.4 Social and Behavioral Sciences

Keywords and phrases narratives, similarity, common summary, empirical study, questionnaire

Digital Object Identifier 10.4230/OASIcs.CMN.2013.129

1 Introduction

Stories play a central role in human knowledge, understanding and reasoning, and are key in explaining human behavior and social communication [1, 24].

According to Schank and Abelson [24], human knowledge is functional, in the sense that all knowledge is encoded as stories along with mechanisms to construct, store and retrieve them. Hence, human memory is a collection of stories that we experienced, heard or composed in the past. Searching in memory is like searching for stories within one's own collection of stories. Moreover, existing stories in memory form our beliefs and are the ingredients of new ideas.

Following that view, understanding and behavior depends upon our past experiences, stored as old stories. In order to understand a new story, we have to find an old story that is *similar* to the new one. Thus, understanding, for a listener, means mapping the speaker's stories onto the listener's stories. This is why different people may interpret the same story in different ways. Analogously, when we face a new unfamiliar situation, we have to determine which of the old situations — stored as stories — in our memory is most relevant to the situation at hand. The familiar situation that is most similar to the new one determines our behavior in the new situation. As a result, intelligent behavior lies upon our capability of finding a relevant past experience that will help us make sense of a new experience [24].

But how does one determine the extent to which stories are considered to be similar?

© Elektra Kypridemou and Loizos Michael;
licensed under Creative Commons License CC-BY
Workshop on Computational Models of Narrative 2013.
Editors: Mark A. Finlayson, Bernhard Fisseni, Benedikt Löwe, and Jan Christoph Meister; pp. 129–146
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1.1 Similarity: Existing Approaches

Similarity is a very powerful construct in psychology, entering into the analysis of many diverse phenomena, such as creative and scientific discovery, problem-solving, categorization, decision-making, learning and transfer [10, 27]. In this section we briefly describe several approaches to similarity as proposed in the literature.

i) The **geometric approach** determines similarity using the *mental distance models* (see, e.g., [20, 25, 26]). According to that view, concepts are represented as points within a multi-dimensional mental space and similarity between concepts as the inverse of the metric distance between these points. Thus, the less the distance between the point representations of two concepts A and B, the more similar the concepts are. As the distance between two points A and B within the space increases, it becomes easier for us to detect that the concepts A and B are different.

The geometric approach treats similarity as a symmetric relation, in the sense that the similarity of A to B equals the similarity of B to A. However, Tversky [27] argued against the symmetry assumption, providing empirical evidence for asymmetric similarities, and proposed his own model for similarity.

ii) The **featural approach**, represented by Tversky's classic *contrast model* [27], proposes that concepts are represented as selections of features, and similarity is described as a feature-matching process. Specifically, the similarity between two concepts is computed based on numbers of shared and not shared features (expressed as a linear combination of the measures of their common and distinct features). The greater the size of the set of the common features of two concepts A and B, and the smaller the sizes of the two sets of distinct features $A \setminus B$ and $B \setminus A$, the greater the similarity between the two concepts A and B. This model also accounts for the asymmetry assumption of similarity, since the negative effects of the two complement sets $A \setminus B$ and $B \setminus A$ are not equal. Hence, if we ask how similar A is to B, then the set $B \setminus A$ counts much more than the set $A \setminus B$.

Later on, Gentner [7] indicated that although Tversky's contrast model seemed to be correct for literal similarity comparisons, it did not provide a good account for other types of comparisons, such as analogies. Hence, Gentner [7] introduced the following approach, providing a theoretical framework for analogy.

iii) The **structural (or relational) approach** is based on Gentner's work on *Structure Mapping Theory (SMT)* [7, 8, 10]. Two concepts, that function as wholes, may be treated as analogous when they share some essential relations, even if they may have a lot of distinct features. Because of this singularity of analogy, structure-mapping, in contrast to the featural approach, treats commonalities and differences as dependent features. Concepts are considered as structures of object attributes and relations, and comparisons are made by mapping the structures of the two concepts. The main idea of this theory is that analogy is characterized by mapping relations between objects, rather than merely mapping the attributes of the objects. Thus, comparisons rely only on the syntax of the knowledge representations and not on their specific content. Furthermore, the structural approach allows analogies to be distinguished clearly from literal similarity and other kinds of concept comparisons, as we shall discuss in more detail in Section 2.2.

iv) A more recent approach, the **tansformational approach** [11, 12] considers concepts as representations, and suggests that similarity depends on the ease of transformation between these representations. Specifically, the less the number of steps needed to transform a representation into another representation, the higher the similarity between the two concepts is. Thus, the more dissimilar the entities, the more transformations are needed. Hahn *et al.* empirically tested the view of similarity as transformational distance [11, 12], and proposed the *Representational Distortion* as a specific example of this approach [11]. However, Larkey and Markman, while testing the similarity judgments for geometric objects, found some evidence against this approach [14].

1.2 Computational Modelling of Narratives

Previous attempts to create computational models of narratives include Propp's narrative functions (or narratemes) [21], Rumelhart's Story Grammars [22], Lehnert's Plot Units [15, 16], Löwe's Doxastic Preference Framework (DPF) [17], Elson's Story Intention Graphs (SIG) [4] and Chambers' Narrative Event Chains (NEC) [2]. Moreover, several recent computational studies provided algorithms that manage to make story comparisons [5, 13] and recognize narrative similarity [18] and analogy [3].

The current study attempts to define similarity of narratives by means of the concept of summary. To this end, it has been conjectured [18, 19] that *similarity between two stories is effectively equivalent to saying that the two stories have a common summary*; i.e., an abstraction that is appropriate for both stories. In particular, the more appropriate this common summary is for the two stories, the more similar the two stories are. This statement forms the main hypothesis that we seek to empirically examine in this work.

2 Background

Our investigation of the above hypothesis does not presuppose any particular framework of story understanding, nor any particular approach to defining similarity. Nonetheless, we find it useful to adopt certain notions and terminology from the SMT [7, 8] and to compare some of our obtained results to those obtained under the SMT. We shall, thus, present the SMT in more detail in this section.

2.1 Structure Mapping Theory for Analogy

When we interpret analogies such as "A is like B", we draw inferences about a concept A (target), based on our knowledge of another concept B (base), which serves as the source of our knowledge. The central idea of the SMT [7, 8] is the definition of the analogy as "an assertion that a relational structure that normally applies in one concept can be applied in another concept" [7]. Thus, the essence of an analogy between two concepts A and B is that they share a common structure. This structure is the dominant aspect of concepts A and B, even though these concepts may differ in many other aspects. For example, we might say that a child's mind is like a sponge. One could easily interpret this analogy, drawing the inference that a child's mind absorbs a lot of information. Hence, we use our knowledge about sponges in order to draw inferences about a child's mind. However, we do not transfer all our knowledge about sponges to the child's mind. If we did so, then we could argue that the child's mind is yellow with holes and holds water. But we do not! Although the child's mind and the sponge do not share other common features, this does not seem to count against the analogy. This is why Gentner criticized Tversky's *contrast model* [27] as not appropriate for analogies [7].

Continuing with the sponge-mind example, it seems that people, seeking to identify analogies, somehow know which features must be transferred and which not. The SMT is capable to explain such behavior, by suggesting that i) we tend to focus on relational information and ignore the distinctive attributes of objects in A and B ii) we prefer to focus

Table 1 Similarity types according to Gentner's work [7, 8, 9].

Similarity type	Object Attributes	Relational Predicates Low-order	Relational Predicates Higher-order
Literal Similarity (LS)	✓	✓	✓
Analogy (ANA)	×	✓	✓
Surface Similarity (SS) or Mere Appearance	✓	✓	×
First Order Relations (FOR)	×	✓	×
Objects Only (OO)	✓	×	×
Anomaly (ANO)	×	×	×

on interconnected systems of relations and favor these relations in our interpretations (see the mapping principles as described later in this section).

According to the structural approach, knowledge is represented as a propositional network of nodes and predicates [7, 23]. Nodes represent concepts, while predicates express propositions about these concepts. Predicates may be either object attributes, taking one argument (e.g., $\text{YELLOW}(x)$), or relationships, taking two or more arguments (e.g., $\text{SMALLER}(x, y)$). A further syntactic distinction of predicates is also made: first-order predicates take objects as arguments (e.g., $\text{HURT}(y, x)$ and $\text{HATE}(x, y)$), while higher-order predicates take propositions as arguments (e.g., $\text{CAUSE}[\text{HURT}(y, x), \text{HATE}(x, y)]$) [7].

Using these notions, we briefly introduce the two main mapping principles of the SMT: i) relations, rather than object attributes, are mapped from the base to the target concept, and ii) a relation that belongs to a system of mutually interconnected relations is more likely to be mapped than an isolate relation (known as the *Systematicity Principle*) [7].

2.2 Similarity Types

The SMT clearly distinguishes analogy from other kinds of concept comparisons. Table 1 summarizes several similarity types, proposed by Gentner *et al.* [7, 8, 9], based on the possible combinations of the predicates that two concepts may share.

To apply Gentner's work to story comparisons, Gentner, Rattermann and Forbus [9] considered i) object attributes as the characters, objects and locations of the story, ii) first-order relations as the events, actions and other relations between the objects of the story (e.g., X talking to Y), and iii) the higher-order relational structure as the causal (or other types of) relations in the story's plot. According to these assumptions, they created pairs of stories by generating several versions of a base story, which differed in the level of similarity they shared with the original story.

In the present study we borrow the names of the similarity types proposed by Gentner (Table 1). However, while applying these types into story pairs, we interpret predicates in a slightly different manner than Gentner *et al.* [9] did (see Section 3.2.2).

3 Empirical Method

In order to test our main hypothesis that story similarity relates to the existence of a common summary, we designed an online questionnaire and we invited people to rate pairs of stories in terms of their similarity or in terms of how appropriate a third story was as a summary of each of the first two stories.

3.1 Participants

Individuals were invited via e-mail to participate voluntarily to the study, by completing an online questionnaire. No financial or other compensation was given for participation. Although more than a hundred people started completing the questionnaire, participants who left before answering at least one question of the main part of the experiment were excluded from our sample. Hence, the final sample of our study comprised 52 adults (21 male, 31 female), aged 18 to 65 years (one reported age 66+). Participants were mainly residents of Greece (26) and Cyprus (22) and they all spoke Greek as their native language.

3.2 Measures

Before proceeding to the detailed description of the various stages of the questionnaire construction, we give a brief overview of the procedure followed, and we introduce the notation that we will use. To prevent any later confusion, we find it useful to discriminate early between the **groups** of stories initially constructed, the **trials** of the questionnaire, and the **triples** of stories used during the analysis stage.

First, we created a pool of selected narratives, serving as the source for our stories (Section 3.2.1). We then edited some of these stories in order to form pairs (A, B) of stories $S \in \{A, B\}$ with varying degrees of similarity (Section 3.2.2). For each story pair (A, B) we also generated 4 other stories $C_{i,j} \in \{C_{11}, C_{10}, C_{01}, C_{00}\}$, which we considered as summaries of stories A and B (Section 3.2.3). As a result, during the story preparation stage, we formed 16 **groups** comprising 6 stories each (A,B,C_{11},C_{10},C_{01}, C_{00}). These groups of stories served as a source for the 16 **trials** of the questionnaire (Section 3.2.4), where each trial comprised 4 stories (A,B,C_{11}, along with one of C_{10},C_{01},C_{00}) selected from its corresponding group. Later on, in order to test our main hypothesis (Section 4.2), for each trial of the questionnaire we formed two **triples** (A,B,$C_{i,j}$), one with $C_{i,j} = C_{11}$ and one with $C_{i,j} \in \{C_{10}, C_{01}, C_{00}\}$, resulting in a total of 32 triples for each participant who fully completed the questionnaire. For participants who completed the questionnaire only partially, we gathered as many triples as possible given the participants' responses.

3.2.1 Selection of a Pool of Original Stories

Since we wanted the stories of the experiment to be as naturalistic as possible and not to appear like artificial stories made especially for use in the lab, we decided to mainly use existing stories, rather than to produce our own. In order to find such original stories, we searched for literature books and collections of myths and fairy tales in the library. We also searched for online videos and texts of narrations, and we recorded oral narrations.

During the selection of the original stories, we mainly took into account two factors: i) the reputation of the story, and ii) the content of the story. Regarding the first factor, to ensure that the participants' previous knowledge of the stories would not bias their responses, we avoided using well-known stories (e.g., the French novella "The little prince" of Antoine de Saint-Exupéry). Concerning the second factor, we gave preference to stories with didactic or entertaining content, which we considered as interesting enough for the participants to read. Having in mind the fact that participation to the questionnaire was a voluntary and time consuming process, such a restriction of the space of different story types was necessary to motivate people to fill in the questionnaire and also to prevent them from leaving the process early. Hence, we avoided stories that merely described a sequence of events without any further purport, and favored those where the author's intention was to pass a deeper message, or even entertain the reader, through the narration. Given these constraints, we created a pool of dozens of original stories, or excerpts of them, both worth reading and not well known.

3.2.2 Generation of Story Pairs According to Similarity Types

Following the story selection, we edited the original stories by adjusting their length. Edited stories were (i) not too long for the participants to read, but also (ii) long enough, in order to keep the style of the original version and not to appear artificial.

We then created story pairs (A,B) with varying degrees of similarity, inspired by the similarity types used in previous work [7, 8, 9] based on the SMT. However, since our aim was to cover the entire range of similarity degrees between stories A and B, rather than to contrast the several similarity types proposed in the literature, we only used Literal Similarity (LS), Analogy (ANA), Surface Similarity (SS) and Anomaly (ANO). Consequently, we defined the variable *sim-type*(A, B), taking values in $\{\mathrm{LS}, \mathrm{ANA}, \mathrm{SS}, \mathrm{ANO}\}$, to indicate the similarity degree assumed by the experimenters for each story pair (A,B).

To create story pairs of the above types of similarity, we interpreted the SMT's predicates (cf. Table 1) as follows: i) object attributes served as the *characters* of the stories, ii) first-order relations served as *relations between the characters*, which could either state a relationship (e.g., X neighbor of Y) or an emotional relation (e.g., X loves Y), while iii) higher-order relations served as the *overall structure of the plot*, in which the relationships and the main interactions between the characters and their actions are represented as an interconnected system (e.g., X lies to Y repeatedly; someday X really needs Y's help; Y do not believe X's need; Y leave X helpless; great disaster happens to X).

Given the above interpretations, we created 4 pairs of stories (A,B) for each of the 4 similarity types, resulting in a total of 16 pairs. To do so we chose a story A from the pool of our selected original stories and we paired it with another story B, which was either generated by the experimenters or, preferably, also extracted from the pool, so as to match the constraints of each similarity type.

Story pairs of the LS type comprised two versions of practically the same story. Hence, we produced story B by expressing story A using different wording, and changing one or two minor details having little or no importance for the overall meaning of the story (e.g., iron wheel vs. stone wheel, for a story where the wheel weight was essential, rather than its material).

Pairs belonging to the ANA type were stories with different characters, where the same basic relations held between them. In the most dissimilar cases of ANA pairs, the overall structures of the two stories differed in some isolated relations, but never in the most interconnected ones.

For example, the structure described earlier (X lies to Y repeatedly; someday X really needs Y's help; Y do not believe X's need; Y leave X helpless; great disaster happens to X) was a common structure of an ANA pair. For story A, X was a shepherd and Y were his fellow villagers, while for story B, X was a grandmother and Y were her children and grandchildren. Some other elements which differed between the stories were: the lie itself, the outcome of lying, the real need of X, and the disaster that happened to X. For example, in story A, the lie was that a wolf attacks the sheep, the outcome was to laugh at the villagers who went to help, the need appeared when a wolf really attacked the sheep, and the disaster was the loss of the sheep. In story B, the lie was that the old woman pretends to be very sick, the outcome was to gain her family's company by 'forcing' them to visit her, the need was when the old woman was really feeling sick, and the disaster was the woman's death.

Table 2 Guidelines used for the construction of the story pairs (A,B) and their summaries $C_{i,j}$ for each similarity type *sim-type*$(A, B) \in \{LS, ANA, SS, ANO\}$.

	(A,B)	C_{11}	C_{10} (resp., C_{01})	C_{00}
LS	Same stories with few minor details differing.	Specific non-shared detail(s) of A and B missing.	C_{11} plus the detail(s) of story A (resp., B).	C_{11} with a key element of the plot changed or missing.
ANA	Stories share a common structure of their plot except for some isolated relations. Different characters.	The common structure of A and B. Characters expressed in an abstract way compatible with both A and B.	C_{11} plus one of the missing isolated relations specified only in A (resp., B).	C_{11} with a key element of the plot missing or replaced to be incompatible with both A and B.
SS	Identical stories up to one point. Different endings.	A summary of the common part of A and B. Ending is missing.	C_{11} plus a summary of the ending of story A (resp., B).	C_{11} plus an ending different than both A's and B's.
ANO	Different stories with a common extremely abstract structure.	The common abstract structure of A and B.	C_{11} plus one of the abstract elements specified in A (resp., B).	C_{11} with one of the elements changed to be incompatible with both A and B.

In order to create story B for the ANA pairs, we searched for stories in the pool and, if we could not find any appropriate story, we generated our own B.

For the SS story pairs, we chose a story A and we generated a story B, identical to A up to some point, but with a completely different ending. The ending was edited in a way that was crucial for the overall meaning of the story (e.g., at the end the king admired and appointed the hero as his main advisor vs. the king disapproved the hero's behavior and banished him: the choice characterizes the whole behavior of the hero as good vs. bad).

Finally, the stories of the ANO pairs were completely different stories that shared some common structure. This common structure was extremely abstract, so that it missed most of the key elements for understanding the overall meaning of the stories (e.g., an older man who lies reveals a secret to a younger man). Concrete characters and relations between the characters (e.g., the secret, the reason the man lies, the reason the young man asks for the secret, and the relation of the two men) differed between the two stories. Similarly to ANA types, we generated our own stories B for the ANO pairs when we could not find any appropriate story B in the pool.

Overall, 20 out of 32 stories of the questionnaire came from the pool of selected original stories (16 as story A, 4 as story B). The remaining B stories were either edited versions of the corresponding original stories A (for the LS and SS types), or stories generated by the experimenters to meet the specific criteria of the ANA and ANO similarity types. Most of the original stories used (14 out of 20) were found in books already translated in Greek, with 6 of them being fairy tales and myths from all over the world (Japan, China, Arabia, Lithuania, Burma, Hungary). Among the remaining original stories, 3 came from online texts, 2 from online videos, and 1 from oral narration.

3.2.3 Generation of Summaries According to Summary Types

For each of the 16 story pairs (A,B) we created 4 different summaries $C_{i,j} \in \{C_{11}, C_{10}, C_{01}, C_{00}\}$, and for each triple ($A$,$B$,$C_{i,j}$,) we defined the variable *sum-type*$(C_{i,j}, S)$, to indicate the appropriateness of summary $C_{i,j}$ for story $S \in \{A, B\}$. Our intention here was to cover the range of possible summaries in a semi-systematic manner. Thus, we created i) a 'common' summary C_{11}, with the aim of being an appropriate summary for both the stories A and B, ii) a summary C_{10}, with the aim of being appropriate for story A and inappropriate for story B, iii) a summary C_{01}, with the aim of being inappropriate for story A and appropriate for story B, and finally, iv) a summary C_{00}, with the aim of being inappropriate for both the stories A and B.

Given the above definition of $C_{i,j}$, the variable *sum-type*$(C_{i,j}, A)$, indicating the appropriateness of summary $C_{i,j}$ for story A, had the value 'good' if $i = 1$, and the value 'bad' if $i = 0$. Accordingly, the variable *sum-type*$(C_{i,j}, B)$, indicating the appropriateness of summary $C_{i,j}$ for story B, had the value 'good' if $j = 1$, and the value 'bad' if $j = 0$. Hence, i can be seen as an index of the appropriateness of summary $C_{i,j}$ for story A, and j as an index of its appropriateness for story B. In general, the variable *sum-type*$(C_{i,j}, S)$ had the value 'good' if either ($S = A$ and $i = 1$) or ($S = B$ and $j = 1$), and the value 'bad' otherwise.

We considered a summary $C_{i,j}$ as inappropriate, when some essential feature of the story S was either missing or changed in the summary. However, appropriate summaries could include an unimportant detail of a story as well. Thus, we did not consider the succinctness of the summary as a defining characteristic of the appropriateness of the summary. Table 2 describes the methodology we adopted in order to create the 4 summaries $C_{i,j}$, for each of the 4 similarity types. C_{11} served as a baseline for creating the other 3 summaries. For C_{10} and C_{01} we modified C_{11} so as to be slightly more appropriate than C_{11} for one of the two stories and slightly less appropriate for the other story, while for C_{00} we modified C_{11} so as to be slightly less appropriate than C_{11} for both A and B. Our concern was to avoid creating summaries that would be *obviously* inappropriate for one of the two stories. This choice was guided by the fact that our main hypothesis states that the existence of an appropriate common summary indicates a high similarity pair. Consequently, low values of summary scores would be less useful in validating or falsifying this hypothesis during our analysis.

3.2.4 Questionnaire Construction

Each trial of the questionnaire included the presentation of 4 stories (A, B, their common summary C_{11}, and one of the non-common summaries), along with 5 questions, asking participants to rate the similarity between stories A and B, and the appropriateness of the other two stories as summaries for each of the stories A and B.

In the place of the non-common summary we chose C_{00} to appear in half of the cases, and C_{10} and C_{01} to appear in one fourth of the cases each. To achieve these frequencies, we created quadruples (C_{10}, C_{01}, C_{00}, C_{00}) and we used the Latin Squares method to counterbalance their order. We created 4 conditions (Table 3) and we randomly allocated participants to each of these conditions (see Section 3.3). As a result, for each of the 4 similarity types, in addition to the common summary C_{11}, each participant was presented with each of summaries C_{10} and C_{01} once, and with summary C_{00} twice. Moreover, the order of the 16 trials of the questionnaire, as well as the order of the two summaries of each trial, was randomized.

Using the open source survey application LimeSurvey 2.00+, we implemented the above methods by constructing an online questionnaire, which we describe in more detail next.

Table 3 Summaries presented across the four Conditions (in addition to C_{11}, which was presented in each Condition).

	Similarity	Condition					Similarity	Condition			
Trial	type	1	2	3	4	Trial	type	1	2	3	4
1	LS	C_{10}	C_{01}	C_{00}	C_{00}	9	SS	C_{10}	C_{01}	C_{00}	C_{00}
2	LS	C_{01}	C_{00}	C_{00}	C_{10}	10	SS	C_{01}	C_{00}	C_{00}	C_{10}
3	LS	C_{00}	C_{00}	C_{10}	C_{01}	11	SS	C_{00}	C_{00}	C_{10}	C_{01}
4	LS	C_{00}	C_{10}	C_{01}	C_{00}	12	SS	C_{00}	C_{10}	C_{01}	C_{00}
5	ANA	C_{10}	C_{01}	C_{00}	C_{00}	13	ANO	C_{10}	C_{01}	C_{00}	C_{00}
6	ANA	C_{01}	C_{00}	C_{00}	C_{10}	14	ANO	C_{01}	C_{00}	C_{00}	C_{10}
7	ANA	C_{00}	C_{00}	C_{10}	C_{01}	15	ANO	C_{00}	C_{00}	C_{10}	C_{01}
8	ANA	C_{00}	C_{10}	C_{01}	C_{00}	16	ANO	C_{00}	C_{10}	C_{01}	C_{00}

3.3 Procedure

An e-mail was sent to a number of people, inviting them to complete the questionnaire and to further disseminate the invitation. In order to encourage participation, in the invitation e-mail we informed the recipients that during completion of our on-line questionnaire they were going to read interesting stories and myths from all over the world. They were also informed that the study concerns the similarity of narratives and summarization, but they were given no further information on the specific hypothesis of our study.

Participants who entered the questionnaire were assured that responses were collected anonymously. After certifying that they are adults and native Greek speakers, they were given navigation instructions, and they were instructed how to safely interrupt the process in case they would wish to continue later. Afterwards, they provided information on their demographic characteristics (gender, age, educational level, country of residence) and they reported whether they enjoy reading stories, myths, fairy tales and literature, and how often they read such texts.

The practice section followed, where an example of a trial was given in order to clarify the process. Participants were told that they are going to read two stories, A and B, and rate their similarity in a scale from 0 (no similarity) to 5 (high similarity). For illustrative purposes, the stories A and B used in the example were considerably shorter than the stories of the experiment. Some help was given, where we stated explicitly the differences between the two stories and asked participants whether they consider these differences (along with others that they may have identified) as significant enough so as to rate the two stories with a low degree of similarity, or they consider the differences meaningless for the general meaning of the stories, and so they would rate the two stories with a high score. After registering their responses, we informed them that two more stories would appear on the screen in order to evaluate how good they consider each of the new stories as a summary for each of the stories A and B, in a scale from 0 (very bad summary) to 5 (very good summary). The stories A and B, along with two example summaries were presented. As previously, some help was provided, stating that some statements of the original stories are missing in the summaries, and asking participants to evaluate whether they consider these statements as important enough so as to rate the summary as bad, or not. Finally, participants were given the chance to read the instructions once more, in case the did not feel confident with the process. At no point in this practice section did we suggest what an appropriate score would be for the example questions.

Following the practice section, a screen with some guidelines appeared. Participants were first encouraged to answer according to their own judgment, since there are no "right" or "wrong" answers. They were then advised to rate summaries by considering which facts of the stories A and B are included and which are missing, and ignore possible syntactic or stylistic aspects of the stories. We also asked participants to be as consistent as possible in their ratings during the whole experiment. Finally we advised them not to judge a summary solely by its length, since length does not necessarily indicate a summary's appropriateness. After reading these guidelines, we asked them to make sure that they are in a quiet environment, since the main part of the experiment was about to start.

At the beginning of the main part of the experiment, each participant was randomly allocated to one of the four conditions of the experiment. Then, the first randomly selected trial started. Participants were blind to the labeling of the story pairs (A, B) according to the 4 similarity types, and the labeling of the summaries $C_{i,j}$ according to the 4 summary types. Hence, they could only see 4 stories presented on the screen, without having any cue about the experimenters' assumptions (i.e., $\textit{sim-type}(A, B)$ and $\textit{sum-type}(C_{i,j}, S)$). The first pair of stories (A, B) appeared and participants were asked to rate their similarity on a scale from 0 (no similarity) to 5 (high similarity). After registering their score, which was recorded as $\textit{sim-score}(A, B)$, two summaries appeared on the screen in random order: the common summary C_{11}, along with one of the summaries C_{10}, C_{01}, C_{00}, according to the trial and the participant's condition (Table 3). Participants were asked to judge how good they considered each of the two summaries for each of the stories A and B, on a scale from 0 (very bad summary) to 5 (very good summary). These ratings were recorded as $\textit{sum-score}(C_{i,j}, A)$ and $\textit{sum-score}(C_{i,j}, B)$ respectively. As a result, for each trial we recorded 1 similarity score and 4 summary scores. After completing the first trial the participants proceeded to the next one and continued scoring as previously, until all 16 trials were completed. Finally, participants were thanked and were invited to leave comments about the survey.

4 Empirical Results

Although only 28 out of 52 participants fully completed the questionnaire, we decided to keep incomplete responses in our analysis, since the experimental design allowed us to do so (the trials of the questionnaire were randomly ordered and independent of one another). For those who fully completed the questionnaire the whole procedure lasted on average 87 minutes. The long duration indicates that participants spent enough time to read the instructions and that they read the stories carefully before they gave their scorings. Table 4 displays the demographic characteristics of our sample. The fact that most of the participants reported that they like reading stories and they do so frequently (see Table 4), indicates that our sample had the appropriate skills and experience for completing the task of our study.

Regarding our experimental design, our data suggests that participants were well distributed over the 4 conditions (12 in the first, 10 in the second, 16 in the third and 14 in the fourth condition). The Latin Squares method also worked efficiently since, in addition to the common summary C_{11} which was presented in all 551 trials, the second summary was chosen 25% (137/551) of the times to be summary C_{10}, another 25% (137/551) of the times to be summary C_{01}, and 50% (277/551) of the times to be summary C_{00} (see 'Overall' row, Table 5).

Table 4 Number of Participants (NOP) for each Age Range, Educational Level, Preference and Frequency of Reading Stories.

Age		Educational Level		Reading Stories		
Range	NOP	Degree	NOP		Like	Do not Like
18–20	2	High School	7	Never	0	1
21–30	19	Bachelor's	24	Rarely	5	6
31–40	9	Master's	6	Sometimes	14	0
41–50	6	Ph.D.	15	Often	20	0
51–60	10			Always	6	0
>60	6					

4.1 Participant Scores Compared to Experiment Type

As a preliminary analysis to validate our experimental setting, we investigated whether the types of similarity and summaries that we had considered, indeed matched the scores given by participants. Participant similarity ($\textit{sim-score}(A, B)$) and summary ($\textit{sum-score}(C_{i,j}, S)$) scores confirmed the experimenters' assumptions regarding the similarity ($\textit{sim-type}(A, B)$) and summary ($\textit{sum-type}(C_{i,j}, S)$) types, respectively. The similarity scores $\textit{sim-score}(A, B)$ not only matched the similarity level assumed for each similarity type $\textit{sim-type}(A, B)$, but also, considering the error bars of Figure 1 (Left), clearly differentiated the LS and ANA (average similarity scores 4.64 and 3.44, respectively) from the SS and ANO types (average similarity scores 2.07 and 1.64, respectively).

Moreover, the 4 summary types $\textit{sum-type}(C_{i,j}, S)$ were also confirmed by the summary scores $\textit{sum-score}(C_{i,j}, A)$ and $\textit{sum-score}(C_{i,j}, B)$ given by the participants for the appropriateness of summary $C_{i,j}$ for stories A and B, respectively (Figure 1, Right). Participants judged the common summary C_{11} as appropriate for both stories A and B. They also judged summary C_{10} as more appropriate for story A than for B, and summary C_{01} as more appropriate for story B than for A. Finally, they judged summary C_{00} with a relatively low degree of appropriateness for both stories A and B (Figure 1, Right; Table 5, 'Overall'). Table 5 shows how the average summary scores $\textit{sum-score}(C_{i,j}, S)$ are distributed across the 4 similarity types. It is noteworthy that in the case of the SS and ANO similarity types, scores given to the common summary C_{11} were considerably lower than the $\textit{sum-score}(C_{10}, A)$ and $\textit{sum-score}(C_{01}, B)$ scores, which is in accordance with our main hypothesis (see Section 5 for a discussion).

4.2 Testing our Main Hypothesis

Figure 2 already provides a first indication in support of our hypothesis, showing that pairs of stories that were of types $\textit{sim-type}(A, B)$ that are considered more similar, were also associated with a better common summary C_{11}, as measured by the participants' average summary score $\textit{sum-score}(C_{i,j}, S)$.

For the main part of our analysis, we ignored the labels given to the similarity and summary types by the experimenters, and we analyzed the relations between the triples of scores given by the participants for each triple of stories (A,B,$C_{i,j}$): i) the appropriateness of summary $C_{i,j}$ for story A ($\textit{sum-score}(C_{i,j}, A)$), ii) the appropriateness of summary $C_{i,j}$ for story B ($\textit{sum-score}(C_{i,j}, B)$), and iii) the similarity between stories A and B ($\textit{sim-score}(A, B)$).

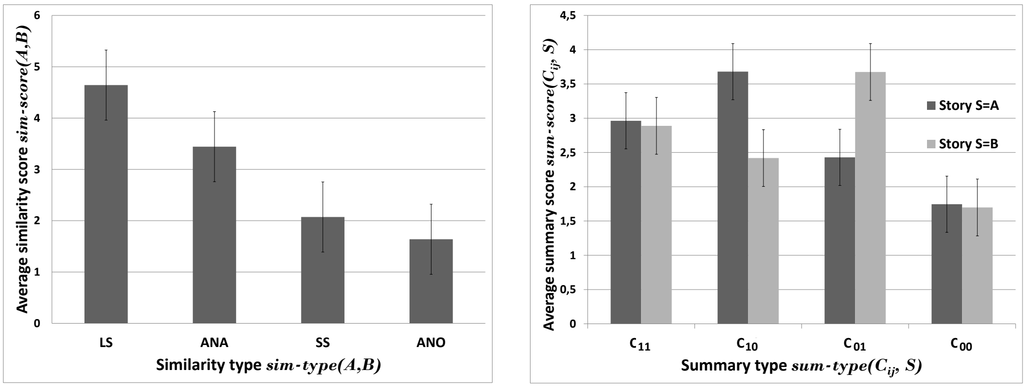

Figure 1 Participants' scorings matching the experimenters' labeling. **Left:** Average similarity scores $sim\text{-}score(A, B)$ given for each similarity type $sim\text{-}type(A, B)$. **Right:** Average summary scores $sum\text{-}score(C_{i,j}, S)$ given for each summary type $sum\text{-}type(C_{i,j}, S)$ with $S \in \{A, B\}$ and $i, j \in \{0, 1\}$. Error bars represent standard errors.

Table 5 Average summary scores $sum\text{-}score(C_{i,j}, S)$ for each Story $S \in \{A, B\}$ and for each summary $C_{i,j}$ with $i, j \in \{0, 1\}$, across the 4 similarity types $sim\text{-}type(A, B)$. Number of responses for each summary $C_{i,j}$ and each story S are given in parentheses.

Similarity type	**Average summary score** $sum\text{-}score(C_{i,j}, S)$ for Story $S = A$, $S = B$ (number of responses)			
$sim\text{-}type(A, B)$	C_{11}	C_{10}	C_{01}	C_{00}
LS	4.39, 4.27 (142)	4.29, 4.15 (34)	3.97, 4.24 (38)	2.53, 2.51 (70)
ANA	3.43, 3.32 (135)	3.43, 2.63 (35)	2.93, 3.61 (28)	2.21, 2.31 (72)
SS	1.99, 1.93 (136)	4.27, 1.45 (33)	1.22, 4.06 (36)	1.25, 1.06 (67)
ANO	2.04, 2.01 (138)	2.69, 1.34 (35)	1.34, 2.66 (35)	0.96, 0.91 (68)
Overall	2.96, 2.89 (551)	3.68, 2.42 (137)	2.43, 3.67 (137)	1.74, 1.70 (277)

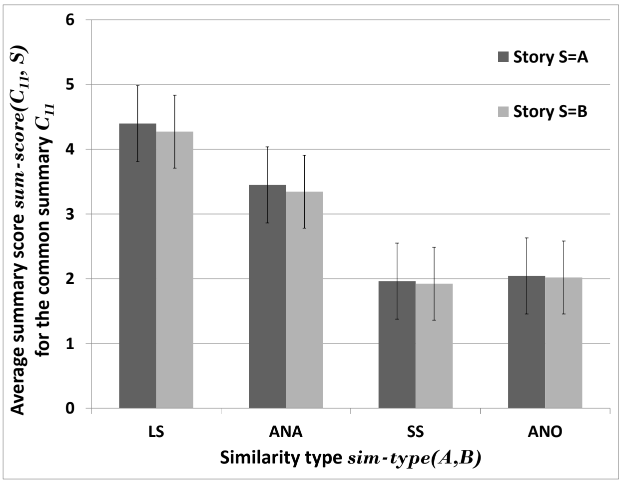

Figure 2 Average summary scores $sum\text{-}score(C_{11}, A)$ and $sum\text{-}score(C_{11}, B)$ for the common summary C_{11} for each of the 4 similarity types $sim\text{-}type(A, B)$. Error bars represent standard errors.

Table 6 Average similarity scores $sim\text{-}score(A, B)$ distributed across the possible summary score pairs $sum\text{-}score(C_{i,j}, A)$ and $sum\text{-}score(C_{i,j}, B)$, for all $C_{i,j}$ with $i, j \in \{0, 1\}$. Similarity scores above 2.50 are shown in bold and similarity scores above 3.00 are underlined.

$sum\text{-}score(C_{i,j}, B)$ \ $sum\text{-}score(C_{i,j}, A)$	*0*	*1*	*2*	*3*	*4*	*5*
5	1.33	2.09	**2.71**	**<u>3.00</u>**	**<u>4.22</u>**	**<u>4.72</u>**
4	2.06	**2.55**	**<u>3.38</u>**	**<u>3.39</u>**	**<u>3.74</u>**	**<u>4.22</u>**
3	1.43	1.60	**2.75**	**2.93**	**<u>3.76</u>**	**<u>3.38</u>**
2	1.88	1.91	**2.98**	**2.81**	**2.78**	**<u>3.89</u>**
1	1.31	**2.52**	1.85	1.71	2.06	1.44
0	1.87	1.36	1.00	1.29	2.00	1.88

Table 6 presents the average similarity scores $sim\text{-}score(A, B)$ given by participants, for each possible combination of scores $sum\text{-}score(C_{i,j}, A)$ and $sum\text{-}score(C_{i,j}, B)$ given for a particular summary $C_{i,j}$. Interpreting the table as a graph, each cell of the table corresponds to a point (x, y) in the graph, with coordinates the two summary scores $sum\text{-}score(C_{i,j}, A)$ and $sum\text{-}score(C_{i,j}, B)$ given for a summary $C_{i,j}$. For example, to determine the value placed in cell (0,0) (i.e., in the lower left corner of the graph), we gathered all the triples (A,B,$C_{i,j}$) for all the summaries $C_{i,j}$ scored with 0 for both stories A and B and we computed the average similarity score $sim\text{-}score(A, B)$ between stories A and B for these triples. Similarly, the average similarity score $sim\text{-}score(A, B)$ for all the summaries $C_{i,j}$ scored with 1 for their appropriateness for story A and with 0 for story B is presented in the cell $(1, 0)$ of the table, and so on. Observe that for triples where summaries $C_{i,j}$ were rated with high summary scores for both stories A and B (upper right corner of the graph), the average $sim\text{-}score(A, B)$ between stories A and B was higher. Hence, the existence of an appropriate common summary for two stories suggests that the two stories are highly similar.

The data of Table 6 is also represented by the contour of Figure 3 (Left), where higher average similarity scores are represented with warmer colors than lower average similarity scores. Accordingly, the 3D curve of Figure 3 (Right) represents the two summary scores $(sum\text{-}score(C_{i,j}, A), sum\text{-}score(C_{i,j}, B))$ by a point (x, y) in the horizontal plane, and the average similarity score for each point (x, y) by the height z of the curve in the vertical axis. Taken together, Table 6 and the two graphs of Figure 3, confirm our main hypothesis: as the summary scores $sum\text{-}score(C_{i,j}, A)$ and $sum\text{-}score(C_{i,j}, B)$ increase, indicating that the summary $C_{i,j}$ becomes more and more appropriate for stories A and B, the average similarity score $sim\text{-}score(A, B)$ also increases.

Finally, we attempted to more systematically measure the appropriateness of a summary $C_{i,j}$ for *both* stories A and B simultaneously, with a single score. Using the two summary scores $sum\text{-}score(C_{i,j}, A)$ and $sum\text{-}score(C_{i,j}, B)$, we estimated the *appropriateness of summary $C_{i,j}$ as a common summary of A and B* with 4 different computational methods: i) the *sum* of scores, ii) the *product* of scores, iii) the *Euclidean (L_2) distance* between scores and the point (5,5), subtracted from the maximum distance between two points (which is $5\sqrt{2}$), and iv) the *minimum* of the two scores. The above functions were selected so that the higher the values they produce, the more appropriate $C_{i,j}$ would be as a *common summary* of stories A and B. When graphing the relation between the average similarity score $sim\text{-}score(A, B)$ and the degree of $C_{i,j}$ as a common summary, as computed using each of the above methods, the relation was found to be directly proportional, and the two quantities extremely highly correlated, confirming our main hypothesis once again (Figure 4).

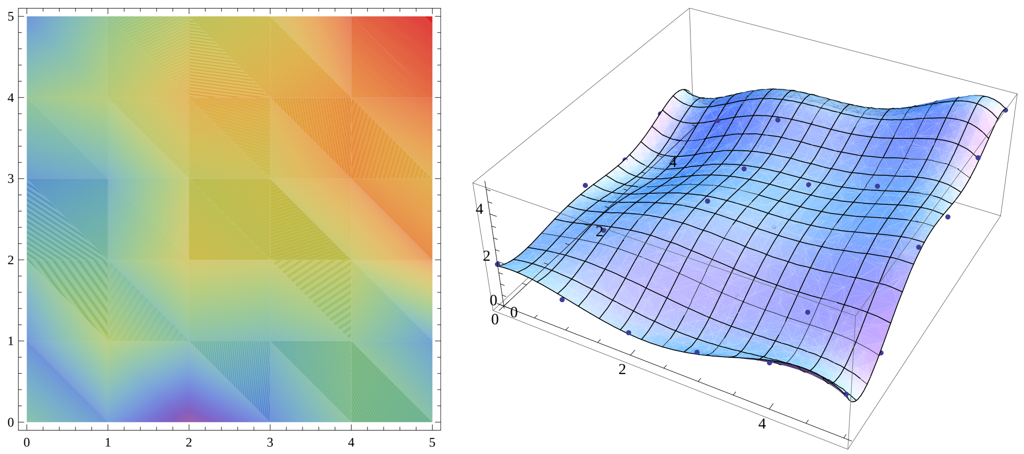

Figure 3 Average similarity scores $\textit{sim-score}(A, B)$ distributed across the possible summary score pairs $\textit{sum-score}(C_{i,j}, A)$ and $\textit{sum-score}(C_{i,j}, B)$, for all $C_{i,j}$ with $i, j \in \{0, 1\}$. **Left:** Contour plot, where average similarity scores are represented by the colors of the spectrum; warmer colors (red) indicate higher similarity scores, while colder colors (violet) indicate lower similarity scores. **Right:** 3D plot, where average similarity scores are represented by the plot height (vertical axis).

5 Discussion

Taken together, the results of our study strongly confirmed our initial hypothesis: stories that are both summarized by a single summary are similar to each other, and the more appropriate this common summary is, the stronger the similarity of the stories becomes. The common summary C_{11} was meant to be an appropriate summary for both stories A and B, and we attempted to choose it to be so during the design of our experiment. However, summary scores given to C_{11}, regarding stories A, B of the SS and ANO similarity types, were relatively low, indicating that this summary was not appropriate enough for the SS and ANO pairs. Accordingly, it seems that we were unable to produce, even though we tried to, an appropriate enough C_{11} summary for the story pairs of low similarity. This is directly in line with our main hypothesis: dissimilar story pairs cannot have appropriate common summaries.

Moreover, the participants' ratings for similarity between the story pairs (A,B) were in accordance with relevant previous work on similarity judgments for stories [9]. Gentner, *et al.* designed several story pairs according to the similarity types of Table 1 and asked participants to rate, among others, the subjective similarity of these pairs on a scale rating from 1 to 5. According to their results, story pairs of the LS type were considered as more similar than those of the ANA type, and story pairs of the ANA type as more similar than those of the SS type. In their study they did not use any story pairs of the ANO type. However, given the definition of the Anomaly type as stories that share none of their predicates, we could plausibly assume that story pairs of the ANO type would be rated with the lowest degree of similarity among all types. Moreover, in Gentner *et al.* [9], the average similarity ratings for the ANA type were close enough to the ratings for the LS type, while ratings for the SS type were much lower. Our results on similarity judgments perfectly reproduced the pattern found in previous work, indicating that the ordering of the similarity types reflects the degree of relational overlap between the pairs.

However, in a previous attempt to empirically determine the factors that affect human

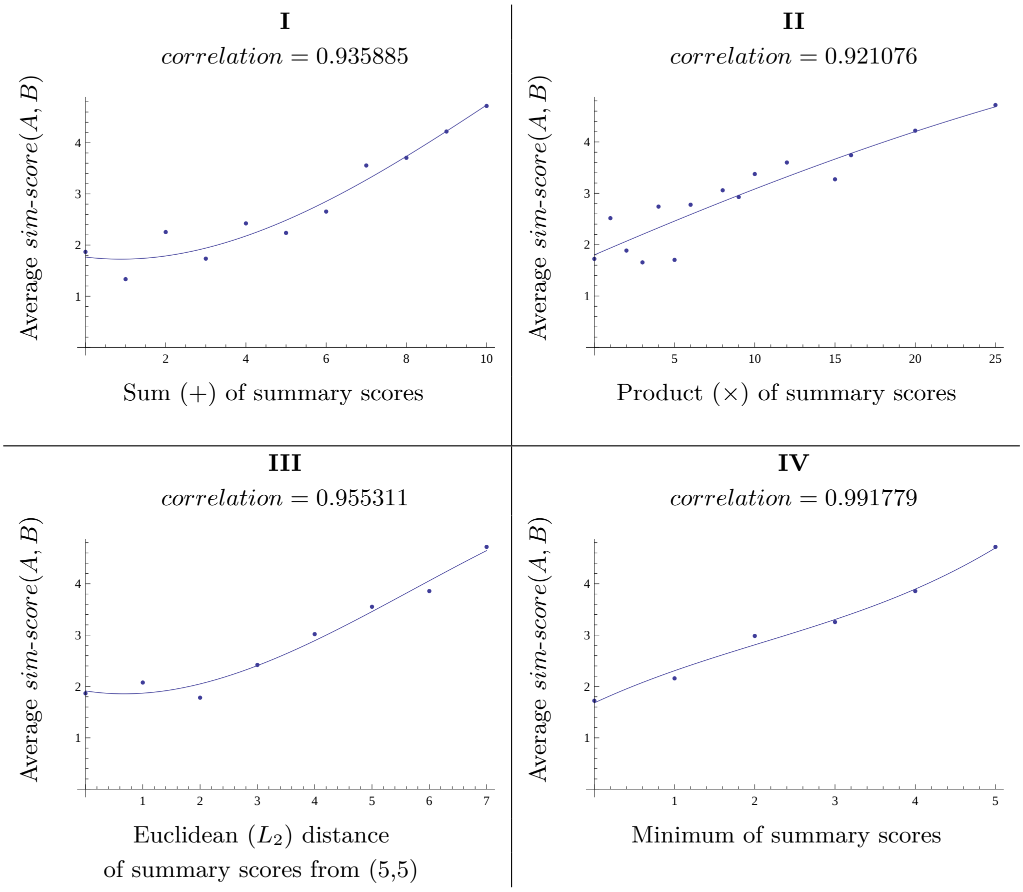

Figure 4 Average similarity score $sim\text{-}score(A, B)$ (y-axis) over the degree of appropriateness of $C_{i,j}$ as a common summary of A and B (x-axis), for each $i, j \in \{0, 1\}$. In each case the x-axis value was calculated as a function F of the two summary scores $sim\text{-}score(C_{i,j}, A)$ and $sim\text{-}score(C_{i,j}, B)$. **I:** $F_I(C_{i,j}, A, B) = sim\text{-}score(C_{i,j}, A) + sim\text{-}score(C_{i,j}, B)$ **II:** $F_{II}(C_{i,j}, A, B) = sim\text{-}score(C_{i,j}, A) \times sim\text{-}score(C_{i,j}, B)$ **III:** $F_{III}(C_{i,j}, A, B) = 5\sqrt{2} - d[(sim\text{-}score(C_{i,j}, A), sim\text{-}score(C_{i,j}, B)), (5, 5)]$ **IV:** $F_{IV}(C_{i,j}, A, B) = min[sim\text{-}score(C_{i,j}, A), sim\text{-}score(C_{i,j}, B)]$

judgments on story comparison, Fisseni and Löwe [6] concluded that structural factors are not the most important aspect for subjective similarity. This discrepancy with our and Gentner *et al.*'s results may be due to the different methodology used. Fisseni and Löwe asked participants to judge pairs of stories as same or different. In a second experiment, participants were additionally asked to justify their sameness ratings by stating as many differences between the stories as possible. This instruction, to explicitly state differences, might have affected their responses. People tend to find more differences for highly similar than for less similar pairs [8]. Since the story pairs used in that study were highly similar (variants of the same story), participants may have reported many differences, even though the stories were similar enough. Accordingly, this may have biased their sameness judgments.

Our empirically tested hypothesis comes from a previous work of Michael [19], as part of a logic-based theoretical framework attempting to computationally define aspects of story understanding. The formal definitions given therein offer qualitative metrics of the appropriateness of a summary and the degree of similarity between two stories. A credulous common summary of two stories is defined as a summary that includes at least the most important parts of the stories; and the existence of a credulous common summary is taken to imply that the two stories are credulously similar in that they share at least their most important parts. The present empirical study, other than to empirically support the psychological validity of the hypothesis offered in that work, can also be seen to extend the logic-based framework by defining some quantitative metrics for the 'credulousness' of a common summary.

The standard distance metrics $L_1 = 10 - F_I$, $L_2 = 5\sqrt{2} - F_{III}$ and $L_\infty = 5 - F_{IV}$, based on the functions F_I, F_{III}, and F_{IV} as previously defined (Figure 4), represent the distance between the points $(\textit{sum-score}(C_{i,j}, A), \textit{sum-score}(C_{i,j}, B))$ and $(5, 5)$. Considering that the point $(5, 5)$ represents the common summary that is the most appropriate for both stories A and B (given that 5 is the maximum score for the score $\textit{sum-score}(C_{i,j}, S)$), it follows that the lower the value of these metrics, the closer the points are to the 'most appropriate common' summary. Borrowing the notion of 'credulousness' of a summary [19], we could name these quantities as "metrics of the credulousness of a common summary".

Finally, we may say that our results reflect the two mapping principles of the SMT. The fact that participants judged story pairs of the ANA type as more similar than those of the SS type, indicates that people consider relational matches as more important than object matches and also, in accordance to the systematicity principle of the SMT, that higher-order relations count more than first-order relations for people's judgments.

6 Conclusion

Identifying similarities among stories is a central part of the process of making sense of stories, and building machines for the latter task will presumably require some solution to the former. In this work we have provided overwhelming psychological evidence that the more appropriate a given story is as a common summary of two other stories, the more similar the latter two stories are to each other. The validity of this hypothesis offers a sufficient condition to test for similarity, or more precisely, offers a way to lower bound the degree of similarity. The condition is not, however, necessary, since the failure of a candidate summary to be an appropriate common summary of two stories does not indicate lack of similarity between the two stories, since some other candidate summary could exist that would be appropriate. Devising a method to produce candidate summaries that would be the most specific common summaries of two stories would offer the missing link to establish the necessity of the condition as well. The role of expectations in stories [18, 19] would seem to be important to that end.

The present study was a first step towards the confirmation of our hypothesis for a certain sample of the possible types of stories. Further research could examine the applicability of this hypothesis to other genres of stories. In a different direction, we could analyze stories extracted automatically from online sources, in order to avoid manually selecting specific types of stories, and any bias this choice may bring to the empirical study.

It would be interesting to generalize our hypothesis beyond stories, and to examine whether similarity between two concepts is effectively equivalent to saying that the two concepts share a common abstraction which is appropriate for both of them. Such concepts could be short videos, simple images, or sound clips. We believe that the empirical methodology developed herein, and the type of analysis performed, could be applied equally well to such more general settings.

References

1 Jerome Bruner. The narrative construction of reality. *Critical Inquiry*, 18(1):1–21, 1991.
2 Nathanael Chambers. *Inducing Event Schemas and their Participants from Unlabeled Text*. PhD thesis, Stanford University, 2011.
3 David K. Elson. Detecting story analogies from annotations of time, action and agency. In Mark Alan Finlayson, editor, *Proceedings of Computational Models of Narrative 2012*, pages 91–99, İstanbul, 2012.
4 David K. Elson. *Modelling Narrative Discourse*. PhD thesis, Columbia University, 2012.
5 Matthew P. Fay. Story comparison via simultaneous matching and alignment. In Mark Alan Finlayson, editor, *Proceedings of Computational Models of Narrative 2012*, pages 100–104, İstanbul, 2012.
6 Bernhard Fisseni and Benedikt Löwe. Which dimensions of narratives are relevant for human judgments of story equivalence? In Mark Alan Finlayson, editor, *Proceedings of Computational Models of Narrative 2012*, pages 114–118, İstanbul, 2012.
7 Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170, 1983.
8 Dedre Gentner and Arthur B. Markman. Structure mapping in analogy and similarity. *Americal Psychologist*, 52(1):45–56, 1997.
9 Dedre Gentner, Mary Jo Rattermann, and Kenneth D. Forbus. The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology*, 25:524–575, 1993.
10 Dedre Gentner and Linsey Smith. Analogical reasoning. In V. S. Ramachandran, editor, *Encyclopedia of Human Behavior*, pages 130–136, Oxford, UK, 2012. Elsevier.
11 Ulrike Hahn, Nick Chater, and Lucy B. Richardson. Similarity as transformation. *Cognition*, 87:1–32, 2003.
12 Ulrike Hahn, James Close, and Markus Graf. Transformation direction influences shape-similarity judgments. *Psychological Science*, 20(4):447–454, 2009.
13 Caryn E. Krakauer and Patrick H. Winston. Story retrieval and comparison using concept patterns. In Mark Alan Finlayson, editor, *Proceedings of Computational Models of Narrative 2012*, pages 119–124, İstanbul, 2012.
14 Levi B. Larkey and Arthur B. Markman. Processes of similarity judgment. *Cognitive Science*, 29:1061–1076, 2005.
15 Wendy G. Lehnert. Plot units: A narrative summarization strategy. In Wendy G. Lehnert and Martin H. Ringle, editors, *Strategies for Natural Language Processing*, pages 375–414, Hillsdale, NJ, 1981. Erlbaum.
16 Wendy G. Lehnert. Plot units and narrative summarization. *Cognitive Science*, 5(4):293–331, 1981.
17 Benedikt Löwe. Comparing formal frameworks of narrative structures. In Mark Alan Finlayson, editor, *Computational Models of Narrative: Papers from the 2010 AAAI Fall Symposium*, number FS-10-04 in AAAI Technical Reports, pages 45–46. AAAI Press, 2010.
18 Loizos Michael. Similarity of narratives. In Mark Alan Finlayson, editor, *Proceedings of Computational Models of Narrative 2012*, pages 105–113, İstanbul, 2012.
19 Loizos Michael. Story understanding... calculemus! In *Proceedings of the 11th International Symposium on Logical Formalizations of Commonsense Reasoning 2013, Ayia Napa, Cyprus*, 2013.
20 Robert M. Nosofsky. Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(1):104–114, 1984.
21 Vladimir Yakovlevich Propp. *Morphology of the Folktale*. University of Texas Press, Austin, TX, 2nd edition, 1968. transl. L. Scott.

22 David E. Rumelhart. *Introduction to Human Information Processing*. John Wiley and Sons, New York, 1977.

23 Roger C. Schank and Robert P. Abelson. *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*. Erlbaum, Oxford, England, 1977.

24 Roger C. Schank and Robert P. Abelson. Knowledge and memory: The real story. In Jr Robert S. Wyer, editor, *Knowledge and Memory: The Real Story*, pages 1–85, Hillsdale, NJ, 1995. Erlbaum.

25 Roger N. Shepard. Stimulus and Response Generalization: A Stochastic Model Relating Generalization to Distance in Psychological Space. *Psychometrika*, 22:325–345, 1957.

26 Roger N. Shepard. Representation of structure in similarity data: Problems and prospects. *Psychometrika*, 39:373–421, 1974.

27 Amos Tversky. Features of similarity. *Psychological Review*, 84:327–352, 1977.