palgrave communications
HUMANITIES | SOCIAL SCIENCES | BUSINESS

ARTICLE

https://doi.org/10.1057/s41599-019-0234-9 OPEN

Cultural entrenchment of folktales is encoded in language

Folgert Karsdorp[1] & Lauren Fonteyn[2]

ABSTRACT In this interdisciplinary study, we explore the understudied effects of growing cultural entrenchment on the form of stories with a long reproduction history. Drawing on insight from literary theory, theoretical linguistics, and cultural evolution theory, we argue that changes in the cultural entrenchment of fairy tales and folk stories are reflected in (small) structural 'mutations' in the story. More specifically, we aim to show that with the increasing familiarity of "Little Red Riding Hood", its story frame and characters have gradually become part of the author and audience's shared world knowledge, which is encoded in the type of linguistic devices used to introduce the characters. To this end, we performed a Bayesian logistic regression analysis on a diachronic collection (late 18th century–2015) of the world's most iconic fairy tale, using automatically generated time estimations for a subset of undated reproductions in the story lineage, and including these estimates and approximated measurement errors in the statistical model. Results show that there is indeed a marked increase of linguistic markers that indicate that the characters are already known or "accessible" to the audience. This development reflects the author's changing intuitions and beliefs about the familiarity of the story, and, indirectly, the changing expectations of the story's audience regarding the appearance of certain characters in the story frame. Notably, this study is the first to quantitatively describe the diachronic development of a story (and the concepts associated with it) into the realm of 'shared knowledge', showing that it is a slow and gradual process. The results help refine our understanding of cultural evolution as well as the workings of speaker-addressee dynamics. Conceptualising the observed linguistic mutations as an instance of guided variation, we argue that the increase of definite first mentions as a function of cultural entrenchment can be treated as an example of variation guided by pragmatic principles such as Grice's Maxim of Quantity, making character introductions as informative as (and not more informative than) required.

[1] Meertens Institute, Royal Netherlands Academy of Arts and Sciences, Amsterdam, Netherlands. [2] English Language and Culture, Leiden University, Leiden, Netherlands. Correspondence and requests for materials should be addressed to F.K. (email: folgert.karsdorp@meertens.knaw.nl)

Introduction

Through their countless reproductions and retellings, fairy tales such as "Snow White", "Little Red Riding Hood" and "Cinderella" have become widely known by Western audiences, and as a result have gradually become deeply and undeniably entrenched in Western culture. In this interdisciplinary study, we aim to explore the understudied effects of growing cultural entrenchment (defined here as 'the actual as well as the assumed/expected knowledge shared between individuals in a cultural community') on the form of stories with a long reproduction history. More specifically, our object of investigation is the world's most iconic fairy tale: "Little Red Riding Hood". The fairy tale's status as the most retold, reinterpreted, recontextualised and reconfigured story in Modern history is virtually unchallenged (Zipes, 1993). As such, "Little Red Riding Hood", with its longstanding reproduction record (and perhaps even longer oral history, cf. Tehrani et al., 2016), is an ideal object for studying whether and how changes in the cultural entrenchment of stories are expressed through and encoded in language.

The current study adds to a growing body of cultural evolutionary research into the history and development of "Little Red Riding Hood" (Tehrani, 2013; Tehrani et al., 2016) and folktales and fairytales in general (e.g., Ross et al., 2013; Stubbersfield and Tehrani, 2013; Da Silva and Tehrani, 2016; Ross and Atkinson, 2016; Bortolini et al., 2017; Acerbi et al., 2017). However, at present, the effect of *entrenchment* on storytelling mainly comes from literary studies, specifically in the context of parody retellings. Here, it is suggested that a reteller's knowledge of their audience's familiarity with the story allows them to play off the audience's associated expectations. In Roald Dahl's famous parodic reinterpretation "Little Red Riding Hood and the Wolf" (1982), for instance, the wolf's attack fails and a well-prepared Red Riding Hood mercilessly shoots him, which is a playful toppling of the outcome expected by the audience. Even with an audience of children, for whom literary devices such as irony and intertextuality are generally deemed less appropriate, it appears that authors *can* and *do* in fact employ such sophisticated narrative strategies, as they "can assume that their young audience is familiar with some version of the tale, however generic it may be, and will read the new story in relation to it" (Beckett, 2002, p. 18). The result of such prior assumptions and expectations of familiarity (or, in other words, the result of cultural entrenchment of a story frame) is thus argued to lie at the core of structural and narrative innovation in reproduction lineages.

Still, while a convincing case can be made that cultural entrenchment plays a pivotal role in the success of parodic retellings, they represent a special (somewhat infrequent) type of retelling, in which substantial (and fully intentional) changes are made to the parodied story. What is perhaps more easily missed are the subtler (and possibly unintentionally used) linguistic elements that signal cultural entrenchment, which are present in a larger if not the entire body of a story's retellings and reproductions. In recent years, a growing number of studies have determined that it is precisely in these more subtle linguistic features that we find various dimensions of cultural, social, conceptual and cognitive information reflected: Louwerse and Zwaan (2009), for instance, have shown that our perception of geographical locations and knowledge about the environment is encoded in language statistics, as are attitudes towards and social relations between persons (Recchia et al., 2014; Hutchinson and Louwerse, 2018). To further address the question whether changes in the cultural entrenchment of fairy tales and folk stories is reflected in (small) structural 'mutations', this study sets out to scrutinise the linguistic encoding of familiarity in a lineage of "Little Red Riding Hood" retellings.

Regarding linguistic encoding of familiarity, defined here as 'shared author-audience knowledge', we are able to build on a rich body of work in language philosophy and theoretical linguistics (see, among many others, Russell, 1905; Christophersen, 1939; Clark and Marshall, 1981; Hawkins, 1978, 1991; Prince, 1992; Givón, 1984; Gundel et al., 1993; Poesio and Vieira, 1998; Epstein, 1999, 2002; Abbott, 2010; Horn and Abbott, 2010). In particular, linguistic encoding of shared knowledge is discussed extensively in the context of how nominal discourse referents (i.e., the people and things we talk about, cf. Sapir, 1921, p. 119) are anchored in the larger discourse (i.e., what is said about them). Generally speaking, speakers/writers (and, by extension, fairy tale authors) have roughly two basic options when they refer to a participant (or character) in a story: either they use a linguistic device that signals to their addressee(s) that they know and are somehow able to identify the person or thing that is discussed (e.g., *I found the book* (*you told me to buy*)), or, alternatively, they use a linguistic device to indicate that a new participant has just been introduced, signalling that the person or thing mentioned is not (yet) known to them (e.g., *I found a book* (*you might like*)). What is interesting in this respect is that it is generally uncontroversial that 'shared knowledge' between a speaker/writer and a hearer/reader is a temporally dynamic entity that can change as the discourse progresses; as soon as a speaker/writer introduces a participant, it is accessible to the addressee and can therefore be marked as 'shared knowledge' at subsequent mention *within* the same text (e.g., *I found a book underneath my old desk. The book is quite old and dusty but you might like it.*). However, there are virtually no empirical investigations into whether the concept of 'shared knowledge' can also be considered a temporally dynamic entity *between* texts, that is, in a reproduction chain of the same story. Our hypothesis, then, is that the effect of increasing cultural entrenchment of a story over time is visible precisely in the changing linguistic encoding of 'shared knowledge' in a reproduction lineage. Put differently, we hypothesise that with the increasing familiarity of "Little Red Riding Hood", its story frame and characters become part of the author and audience's shared world knowledge, which is reflected in the type of linguistic devices used to introduce the characters into the story.

The present study will set out to test this hypothesis by means of a quantitative investigation of a diachronic corpus of Dutch "Little Red Riding Hood" retellings (Karsdorp, 2016). First, in Section 'Indefinite and definite first mention of characters in narration', we provide a more detailed description of the grammatical devices used to encode shared knowledge of referents, drawing on theoretical linguistic descriptions of the concept of *definiteness* and its function in narrative structures. Having provided a more in-depth discussion of the linguistic features under investigation, we turn to a description of the data collection and corpus construction in Section 'Data collection'. Subsequently, Section 'Methodology' presents a description of the predicted and predictor values used in the statistical model adopted in this study, followed by a statistical model specification. After summarising our main results in Section 'Results', the final section of this study offers a discussion of the observed diachronic changes. First, we consider the validity of our findings in light of earlier studies on the oral history and cultural evolution of fairy tales (Tehrani, 2013; Tehrani et al., 2016). Subsequently, we interpret our findings in light of principles and insights gained from theories of pragmatics as well as cultural evolution theory. We conclude with a number of suggestions to further pursue the study of the diachronic development of 'shared knowledge' from a multidisciplinary angle, in which insights from theoretical linguistics are put forward as explanatory factors in the evolution of cultural artefacts constructed with language.

Indefinite and definite first mention of characters in narration

To refer to a character in a larger narrative, writers and storytellers have two basic linguistic strategies to their disposal: they can either present the character (or 'referent') by means of (i) *indefinite* reference (using an indefinite article, e.g., English *a*, Dutch *een*) or (ii) *definite* reference (using a proper name, e.g., *Queen Elisabeth*, or a definite article, e.g., English *the*, Dutch *de*). Indefinite reference is generally used to introduce 'new information', i.e., information the writer deems unidentifiable to the reader, whereas definite reference is typically used for 'given information', i.e., cases where the writer assumes that the reader is able to identify the intended referent (for Dutch, see e.g., Broekhuis and Keizer, 2012, p. 446)[1]. In a typical narrative, the first mention or introduction of a character often coincides with the introduction of new information. Consider the example below, which is a Dutch translation of a fable by Aesop.

1. **Een hond** en **een haan** werden grote vrienden en besloten samen op reis te gaan. Bij het vallen van de avond vloog **de haan** op een tak van een boom en **de hond** kroop in een holte van de stam. Bij het morgenrood werd de haan wakker en kraaide als gewoonlijk. Dat hoorde **een vos** en daar hij ontbijten wilde ging hij onder de boom staan en verzocht de haan beneden te komen: 'Ik zou zo graag kennis maken met iemand die zo'n mooie stem heeft,' zei hij. De haan antwoordde: 'Wil je dan even de portier roepen die aan de voet van de boom slaapt? Dan kan hij je binnenlaten.' Dus klopte **de vos** op de stam, de hond vloog te voorschijn en verscheurde hem. (Arthur Van Schendel, 1978)

   'A dog and a rooster became good friends and decided to go on a trip together. At nightfall, the rooster flew to a branch of a tree and the dog crawled into a hole in the trunk. In the early morning the rooster woke up and crowed as usual. That was heard by a fox, and seeing he wanted breakfast, he went and stood under the tree and asked the rooster to come down. 'I really want to meet with someone who has such a beautiful voice,' he said. The rooster responded: 'Then would you mind calling the porter who is sleeping at the foot of this tree? Then he can let you in.' So the fox knocked on the tree trunk, and the dog appeared and devoured him.'

The short and simple narration set out in (1) consists of an interaction between three characters: a dog, a rooster and a fox. Assuming that the audience is not familiar with the fable and its characters, the author presents its three talking animal protagonists as 'new information' by using indefinite reference at first mention of each character (*een hond* and *een haan* in line 1, *een vos* in line 4). As soon as the protagonists have been introduced, the author can assume they have become identifiable or 'accessible' to the addressee, and thus, they can be *anaphorically* referred to with definite reference at second (and all subsequent) mention (*de haan, de hond* in line 2 and *de vos* in line 7; cf. Hawkins, 1991, p. 406). Thus, the choice between definite and indefinite reference in one way is, as Givón (1984, p. 459) puts it, "a profoundly pragmatic affair", in the sense that a speaker/writer's choice to present a character as an indefinite or definite referent has to do with how the speaker/writer assesses the "current state of knowledge" of their audience at a given point in the narration.

Folktale characters are typically introduced as 'new' at first mention; hence, their referent is introduced in the narrative by means of an indefinite article (e.g., 'Once upon a time there was *a* little village girl…' in Perrault's "Le petit Chaperon rouge", 1697). It should be noted, however, that first mention of a character or 'referent' can also be done by means of definite reference (cf. Poesio and Vieira, 1998). In such cases, definite reference is used not to signal that the referent has previously been mentioned in the narrative (as is the case with anaphoric reference in (1)), but that the intended referent is part of the speaker/writer and hearer/readers shared (extra-textual) socio-cultural knowledge (Givón, 1984, p. 399)[2]. In other words, the speaker/writer can use definite reference to signal to their audience that they are somehow able to identify the intended referent, for instance, when it can be assumed that the referent is a unique, widely-known person, object or concept in a particular culture. The phenomenon whereby definite reference at first mention is warranted by shared cultural knowledge is most commonly illustrated by examples such as the ones illustrated in (2a) and (2b):

(2) a. *Queen Elisabeth* wore an extravagant hat.
    b. *The president* is elected every four years.

The example in (2a) serves to illustrate that, given the right socio-cultural context, a proper name *Queen Elisabeth* can be used to effectively help single out the intended referent to a hearer/reader: in present-day Western Culture, the proper name *Queen Elisabeth* will—even at first mention of the referent—most likely be unproblematically interpreted as the present queen of England (Elisabeth II). Note that proper names "present themselves as being associated with a single referent" and thus possess what is called "referential uniqueness" (Abbott, 2010, Section 2.2.3.1). Speakers/writers using a proper name can expect the addressee to determine who or what is being discussed, but only if the name and its referent (or 'value') are already familiar to the addressee (cf. Prince, 1992, p. 302). The more 'widespread' the knowledge of the proper-name-plus-referent combination in a particular culture or community, the more likely it is that the addressee will successfully identify the referent without (prior) introduction or description (cf. 'larger situation uses' in Poesio and Vieira (1998) and Hawkins (1991)). The example in (2b) is slightly different from the one in (2a) in the sense that, rather than successfully narrowing down the referent to a specific person (or 'value'), the definite noun phrase *the president* refers to a particular 'role' which can be considered part of widely-shared current socio-cultural knowledge.

What is shared between the example in (2a) and (2b) is that the 'first-mention definiteness' of *Queen Elisabeth* and *the president* "need not be justified by a link to a specific frame" (Epstein, 1999, p. 56). There are, however, cases where the appropriateness of definite first mention does depend on a link to a (stereotypical) semantic or narrative frame. For instance, Epstein (1999, pp. 58–59) uses an example of a movie review of *Genesis* (release date 1986), of which the general storyline mimics the original biblical story describing how a woman's arrival marks the beginning of the fall of man:

(3) The film's setting and the story both have a mythic simplicity. In the aftermath of a drought that leaves most people surviving by selling themselves into lifelong servitude, a farmer and a weaver escape and set up residence in a desert ghost town. (…) Then **the woman** arrives, like a fleeing animal. (…) And so begins the slow spiral towards a disaster as ineluctable, no doubt, as the eternal cycles of drought and flood. (Epstein, 1999, p. 59>*Spectator*, Raghleigh, North Carolina; 14/02/199, pp. 11–12)

In this example, Epstein (1999, p. 59) explains, "the author chooses the definite article in order to introduce the concept of 'woman' as a role in what we might call a 'creation story frame'". The use of definite reference here can thus best be understood as a kind of indirect anaphor: a role or character is presented as familiar because it plays a role in a *stereotypical narrative frame* that is part of widely shared socio-cultural knowledge.

It should be uncontroversial to posit that, similar to the creation story, well-known folk and fairy tales also constitute stereotypical narrative frames with more or less fixed unique roles (e.g., hero, girl, stepmother, forest, wolf, villain). This stereotypicality is evidenced by the fact that audiences appear to have expectations about the course of the story and appearance of certain characters: indicative of their expectations, readers often act surprised when they discover that there is no rescuer in Perrault's version of "Little Red Riding Hood". It could even be argued that the entrenchment of roles and characters extends beyond the narrative frame of their particular tale into a larger semantic frame constituted by the body of folk and fairy tales of a particular culture, as suggested by an interesting example provided by Beckett (2002, p. 277), who discusses a range of so-called "fairy-tale medleys":

(4) I was playing with the traditional beginning 'Once upon a time there were three bears…' and I thought, 'Why stop there?' So I added *the seven dwarfs*, and threw in *six gorillas*, *three firemen*, etc." (italics added; Allen Ahlberg cited in Beckett, 2002, p. 277)

In this example, British author Allen Ahlberg describes the thought process underlying one of his own fairy tale medleys, "Jeremiah in the dark woods". Interestingly, his story is not an adaptation of "Snow White and the seven dwarfs", but still he can felicitously use definite reference to refer to the seven dwarfs at first mention simply because they reside in the larger semantic domain of folk and fairy tales. This can be contrasted with the introduction of unconventional fairy tale characters (i.e., *six gorillas, three firemen*), which are introduced by means of indefinite bare plural noun phrases.

What emerges from the discussion of first mention definiteness presented here is that, regardless of whether the locus of connection is a specific narrative frame ('the "Little Red Riding Hood" frame') or a more general semantic frame, the condition that makes definite first mention of the characters possible is that the story frame as well as its characters can be assumed to have reached a certain degree of entrenchment in the socio-cultural knowledge of the speaker/writer and their audience. Our hypothesis, then, is that with the increasing familiarity of the story over time and its consequent entrenchment in culture, the characters of "Little Red Riding Hood" became a more established part of the shared world knowledge of the author and their audience. In other words, as time progresses, we should be able to attest a shift of a story frame and its roles/characters from 'new information' to 'shared cultural knowledge'.

Data collection

The *Koninklijke Bibliotheek* of the Netherlands (National Library) is in possession of a tremendously rich collection of children's books. It consists of over 195 thousand books that have been collected over a period of two hundred years. The collection contains about 630 versions of "Little Red Riding Hood", the oldest version dating from the late 18th century and the latest from 2015. These versions take the shape of a wide variety of visual and written genres, including but not limited to books, picture books, picture stories, colouring pictures, picture postcards, peep shows, catch pennies, ABC books, pop-up books, poems, puzzles, comics, plays, and scripts. Many versions are part of the Special Collections department of the National Library, which contains books and manuscripts that are too old, rare, precious or fragile to be made available through general circulation.

For this study, we required a full-text version of the collection, yet only a handful of retellings of "Little Red Riding Hood" have been made digitally available. As such, corpus construction required all available versions listed in the catalogue of the National Library (with the exception of reprints) to be digitised. Where possible, these versions were scanned, transcribed by Optical Character Recognition (OCR) and manually post-corrected. Additionally, as many versions are part of the Special Collections department, the remaining reproductions were manually digitised. All manual transcriptions were checked by a second transcriber. For reasons of historical accuracy, errors in the original texts, spelling or otherwise, were not corrected.

After removing duplicates, the total number of stories in the digitised version of the collection amounts to 374. To produce a machine-readable edition of the collection, all versions were encoded following the guidelines provided by the Text Encoding Initiative (TEI), which maintains a standard for the representation of texts in digital form. Each story in the collection was annotated with the following metadata fields: (i) title, (ii) author, (iii) publisher, (iv) (estimated) year of publication and (v) ISBN number. These metadata were extracted from the bibliographical records of the stories as provided by the National Library of the Netherlands.[3]

Methodology

The scope of the present study is limited to the introduction of "Little Red Riding Hood"'s two main characters: Red Riding Hood and the wolf. As we hypothesise that the linguistic strategies employed to introduce these characters changes as the cultural entrenchment of the story increased over time, we will test whether there is an effect of time on increased use of definite reference by means of logistic regression analyses. A separate model is built for each character (i.e., one for Red Riding Hood and one for the wolf). Before turning to the description of the logistic regression models, we will briefly discuss the annotation procedure. For each story in the collection, we annotated whether Red Riding Hood and the wolf were introduced at first mention in the story by means of either definite or indefinite reference (i.e., predicted or dependent variable, discussed in Section 'Dependent variable: (in)definite reference'). In Section 'Predictors' we will describe the predictor variables, that is, *Pictorial presence*, *Opening Phrases* and finally, of course, *Time*.

Dependent variable: (in)definite reference. While seemingly straightforward, a few clarifications regarding the annotation procedure of the dependent variable, i.e., linguistic coding of (in) definiteness, should be made. As indicated in Section 'Indefinite and definite first mention of characters in narration', the conventional linguistic device to introduce a character that is assumed to be unfamiliar to the addressee is an indefinite NP (e.g., *een meisje* 'a girl', *een wolf* 'a wolf'), whereas familiar referents (or referents that are uniquely identifiable in a particular context) can be introduced by means of a definite NP or a proper name (e.g., *de wolf* 'the wolf', *Roodkapje* 'Red Riding Hood') at first mention. However, Hawkins (1978) identifies a number of 'unfamiliar uses' of definite first mention where definite coding is triggered various types of modifying elements (also see Poesio and Vieira, 1998):

(5) a. **Nominal modifier**: *De kleur blauw* werkt kalmerend. (*Een kleur blauw)
'The colour blue has a calming effect.'
b. **Relative clause**: Ik word gek van het liedje dat mijn bovenbuurvrouw de hele dag speelt op haar piano. (?een liedje)
'I'm annoyed by the song my upstairs neighbour plays on her piano all day long.'

Table 1 Number of definite and indefinite first mentions of Little Red Riding Hood and the wolf

	RRH	wolf	Total
Definite reference	85	164	249
Indefinite reference	286	204	490
Total	371	368	739

Table 2 Number of stories including pictures of little red riding hood and/or the wolf

	RRH	wolf
Pictured	251	195
Not pictured	123	179
Total	374	374

Table 3 Number of definite and indefinite first mentions of Little Red Riding Hood in stories with traditional and non-traditional opening phrases

	traditional	non-traditional	Total
Definite reference	0	85	85
Indefinite reference	224	62	286
Total	224	150	374

Our data set did not contain any examples where definite coding is triggered by means of pre- or post-modification in the referring noun phrase, as the characters, at first mention, are all introduced by simplex NPs and proper names. However, we did find a number of cases where first mention of the 'unfamiliar' character is not done by means of an indefinite NP. More precisely, what we observe is that Roodkapje, the lead character, is first mentioned by means of a proper name in a copular clause, as in Example (6):

(6) Roodkapje was een kleine meid,
Die door een stout te wezen,
Haar moeder droefheid heeft bereid,
Zooals ge zelf zult lezen.

'Red Riding Hood was a little girl,
Who by being bad
Had made her mother sad
As you will read yourself.'

The issue with examples such as (6) is that, even though the first mention of the character is realised by means of a 'definite' strategy (the subject of the sentence is the proper name *Roodkapje*), the character itself still 'requires' further description and specification (in the subject complement: *een kleine meid, die …*). It appears, therefore, that the referent is not portrayed as 'identifiable' to the addressee, and consequently, these examples were analysed as cases of indefinite reference (in line with how the nominal subject complement is coded in the sentence). Table 1 provides an overview of the raw frequency of definite and indefinite first mentions of Red Riding Hood (RRH) and the Wolf.

Predictors

Pictorial reference. In addition to the factors described in Section 'Indefinite and definite first mention of characters in narration', which mainly dealt with the use of (in)definite reference in the context of new and given or 'accessible' information (either through previous mention or through 'shared world knowledge'), the effect of common ground on the usage of (in)definite reference has been studied extensively in both theoretical and experimental (psycho)linguistic literature (see e.g., Schmerse et al., 2015 and references cited therein). Importantly, studies like Kail and Hickmann (1992) show that children are more likely to use definite reference to introduce characters to a story when they and their interlocutor were looking at a picture book together, allowing them to assume mutual, extra-textually given knowledge (cf. Hickmann et al., 1995 and Schneider and Dubé, 1997). Naturally, experimental results such as these raise the question of whether any developments in the use of definite reference in Red Riding Hood is possibly an effect of a story's accompanying pictures of those characters. That is to say, a picture of the wolf, for example, could function as a deictic point of reference or common ground, allowing subsequent definite first mention.[4] The prominence of the matter is underscored further by the observation that, while pictures have always been common in retellings of "Red Riding Hood", it is especially the 20th century that has seen a strong rise in the number of picture books. Thus, to avoid verifying the hypothesis that the increasing familiarity of the story is reflected in the use of definite reference when it is simply an epiphenomenon of the increasing number of pictures in retellings of the story, we added pictorial presence as a factor to the model. To this end, we annotated all stories in the collection for the presence of a picture of the wolf and a picture of Red Riding Hood. There are 195 stories with and 179 without a picture of the wolf, against 251 stories with a picture of Red Riding Hood and 123 without (Table 2).

Opening phrases. A second necessary measure is the inclusion of fixed, 'traditional' opening phrases. Genre-specific opening phrases, such as "*Long, long ago…*", "*In a country far, far away…*", and "*Once upon a time there was…*", constrain the way in which characters can be introduced to the story, as they commonly include so-called *there*-existential structures that require the use of indefinite reference (the "definiteness restriction", see Milsark, 1974). With Red Riding Hood often being introduced in the opening sentence, it is likely that her introduction is affected by these opening constraints. To quantitatively assess the effect of opening phrases on the choice of indefinite and definite first mention of characters, we include a second predictor and encode each story for containing a traditional opening phrase. In total, 224 stories open with a traditional phrase (as in (7)), against 150 that start using an alternative opening (e.g., in *medias res*, as in (8)).

(7) Alle mooie sprookjes beginnen met "Er was eens…". Dus er was eens een meisje dat in een klein dorpje woonde ergens op het platteland.
'All beautiful fairy tales start with "Once there was…". Thus, once there was a girl who lived in a small village somewhere in the countryside.'

(8) Kijk, daar gaat Roodkapje. Bij elke stap die ze zet, wipt haar rode kapje boven de struiken uit die langs het bospad staan. Bospad?
Ja, ze loopt door het bos naar oma. 'Look, there goes Red Riding Hood. With every step she takes, her red hood teeters above the bushes along the road in the woods. In the woods? Yes, she is walking through the woods to her grandmother.'

Table 3 provides a cross-tabulation of opening types (i.e., traditional versus non-traditional) against definite and indefinite introductions of Red Riding Hood. The statistics highlight that

the presence of traditional opening acts as deterministic factor for the introduction of Red Riding Hood: Red Riding Hood is introduced with indefinite reference in all traditional openings. Such deterministic factors can cause a problem (known as the quasi-separation problem) for logistic regression models in terms of model convergence, as well as unreasonable model and error estimates. Section 'Statistical model specification' discusses how we will deal with this issue.

Time. Finally, to measure the effect of time on the choice for definite or indefinite first mentions of characters, we include the year of publication as a final predictor. Unlike the other two predictors, the inclusion of time as a factor requires virtually no motivation, as it is precisely the effect of increasing cultural entrenchment over time on the occurrence of definite introduction that we wish to investigate. However, what does require some explanation is the particular procedure of how we approached the annotation of *Time*.

As a first step, the dates of publication were extracted from the bibliographical information provided by the National Library. While the majority of stories are provided with exact dates of publication (68%), we have to rely on rough estimates for a smaller yet significant number of stories (32%). These estimates are generally based on common metadata-based assignment methods (e.g., the period in which a publisher or author was active), or rudimentary content analyses (e.g., the examination of certain spelling conventions). In the best-case scenario, these estimates are limited to a particular year. However, often they refer to a timespan of a decade, and sometimes to half or even an entire century. Naturally, this constitutes a relatively large error margin, which poses a problem for investigating the effect of time on the diachronic development of first mention reference. Thus, as a second step, we could turn to essentially three solutions to this problem: (i) we ignore the error margin and pretend the estimates to represent exact publication years; (ii) we discard all stories with publication date estimates; (iii) we embrace the uncertainty of the error margin in our analyses by treating date estimates as dates with *measurement errors*. While the second solution is certainly more principled than the first, it is still the case that inexact estimates of publication dates provide important and valuable information about their true location, leaving the third solution as the most desirable one. The challenge, then, is to adequately embrace and incorporate our knowledge about the error with which the estimates were made.

Following Blackwell et al. (2017), we can think of a publication date w_i as the combination of a true latent value x_i^* and some measurement error u_i, i.e., $w_i = x_i^* + u_i$. There is no measurement error ($u_i = 0$) when a publication date is known. It follows that, for date *estimates*, the error is greater than zero but smaller than infinity ($0 < u_i < \infty$). Finally, if information about the date of publication is entirely absent, the measurement error approaches infinity, leaving us without any information about the true latent date x_i^* (i.e., missing data). What we aim to do is to approximate the error with which publication date estimates were made in order to incorporate the error in the model as well as to make more robust predictions regarding the effect of time. Unfortunately, we cannot simply rely on the publication date estimates provided by the bibliographic records of the story collection because it is unclear what their degree of measurement error is.

A promising remedy to such unclarity is provided by Blackwell et al. (2017), who propose a framework to estimate x_i^* and u_i based on error variance and observed covariates. Here, we take a similar approach, as we opted to re-estimate the publication dates using a content-based procedure. Automatically estimating the publication date of texts has received considerable attention in the literature (see e.g., De Jong et al., 2005; Kumar et al., 2011; Garcia-Fernandez et al., 2011; Bamman et al., 2017; Kulkarni et al., 2018). We adopt the approach suggested by Bamman et al. (2017), who treat publication date estimation as a linear regression problem, with the year of publication as the scalar response variable, and textual features as explanatory predictors. Regarding the textual features, to cope with as well as to maximally exploit any temporally bounded spelling variation in the collection, we employ character-level n-gram features with $1 \leq n \leq 4$. Textual feature counts are sublinearly scaled by taking their log and adding 1, and, subsequently, normalised using ℓ_1 norm. Subsequently, features occurring in less than two documents are excluded. With the remaining features, we train a linear regression model with ℓ_1 and ℓ_2 regularisation (Elastic net regression) on our set of 255 stories with *known* publication years. All parameters (of feature selection and the model) are optimised using a stratified 5-fold cross-validation setup. The performance of our setup is evaluated using the coefficient of determination, R^2, as well as the Mean Absolute Error (MAE) between the estimated years of publication $\widehat{y}$ and their corresponding true years y.

The upper graph in Fig. 1a) displays the aggregated results of performing a 5-fold cross-validation setup on the training data, with the estimated years of publication on the Y-axis against their true values in the training data on the X-axis. The lower graph visualises the residuals. Ideally, the residuals are centred around zero, indicating a clear linear relationship in the data. The model performs well, with an R^2 score of 0.88 and a MAE score of 10.8 with a standard deviation $\sigma = 9.1$. Note that a simple baseline model which always predicts the expected (i.e., mean) year in a training fold regardless of the input features would yield a MAE score of 33.6 (and, by definition, an R^2 score of 0). Adding content-based features thus greatly improves the accuracy of estimation. With confidence in the performance of the model as well as knowledge about its degree of mismeasurement, we can employ it to provide the stories with *uncertain* publication dates with a re-estimated year of publication. Figure 1b) displays the model re-estimations against the date estimations provided in the bibliographic records.[5] Interestingly, the clear linear relationship suggests a global agreement between the re-estimations and estimated bibliographic record dates, but the lower R^2 score (0.73) and higher MAE (14.3) of the correlation model in Fig. 1b) indicates disagreement at local estimations.

While not perfect, a re-estimated publication date w_i can still hold valuable information for our subsequent analysis, if we consider the estimated degree of mismeasurement ($\sigma = 9.1$). Assuming that the degree of measurement error u_i is sampled from, for example, a Normal distribution with mean zero and scale σ, we can include not one but n estimated publication dates by repeatedly sampling a measurement error u_i and subtracting it from the model estimation w_i. Thus, we can effectively construct n parallel datasets with slightly different date estimates, which can subsequently be included in the logistic regression model discussed in Section 'Statistical model specification' below. Note that stories with known, unambiguous publication dates remain unaltered over the parallel data sets, as their measurement error is assumed to be zero. In the analysis below, we fit n logistic regression models to these datasets separately and subsequently pool the results across models. By means of illustration, Fig. 2 displays $n = 1000$ overlaid density plots of the estimated publication years of the stories. The graph clearly shows that

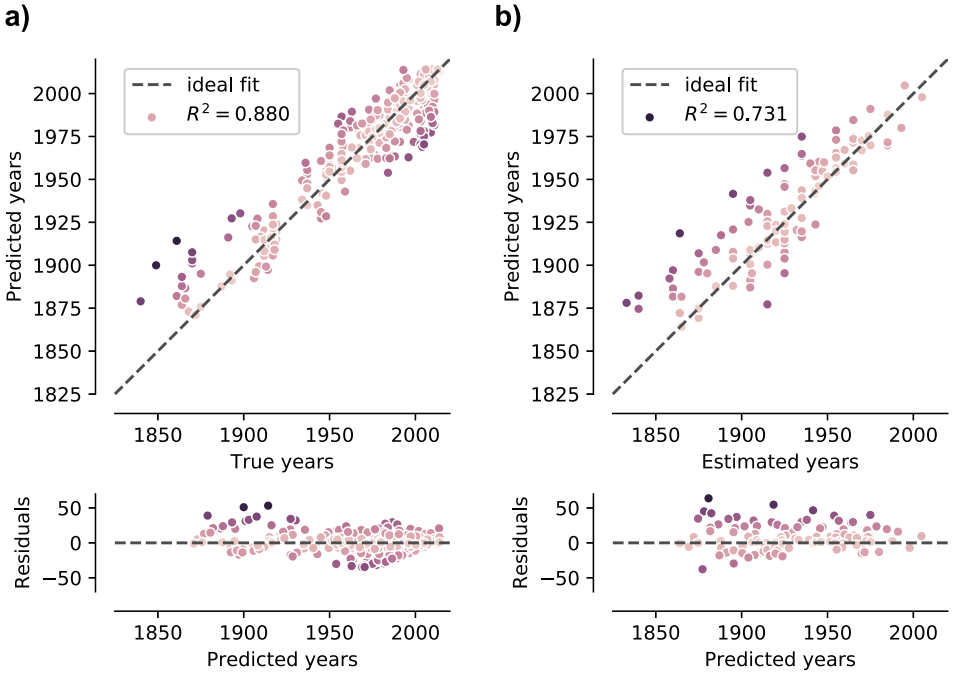

Fig. 1 Results from the date regression experiments. **a** Aggregated results of performing 5-fold stratified cross-validation. The upper graph displays the predicted dates of publication against their true values. The lower graph shows the residuals of the regression. **b** Predictions for all stories with unknown or estimated dates of publication. The graphs display the model estimates against the bibliographical estimates

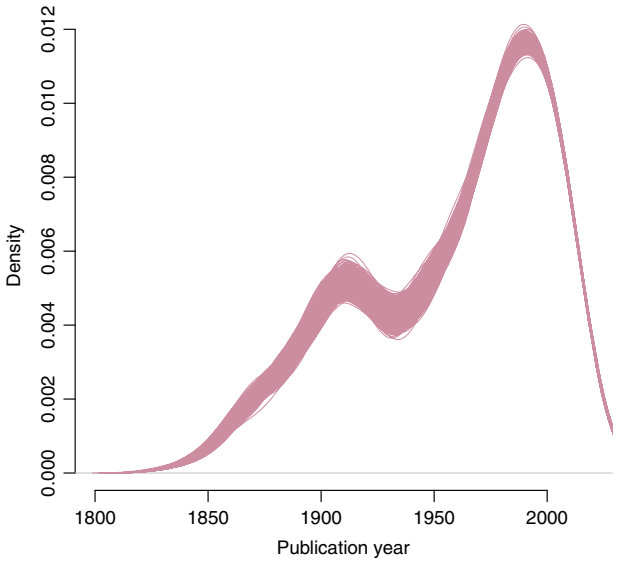

Fig. 2 Overlaid density plots of estimated publication years

most of the uncertainty regarding publication date is situated around the year 1900.

Statistical model specification. We employ Bayesian logistic regression models as implemented in the R package *brms* (Bürkner, 2017)[6]. The 'No U-Turn Sampler' (NUTS) was used for sampling (Hoffman and Gelman, 2014), which is a specific type of Hamiltonian Monte Carlo (cf. Neal, 2011). The choice for Bayesian rather than frequentist methods was motivated by three considerations. First, Bayesian methods often offer superior handling of small sample sizes, which is of crucial importance given the relatively few data points in our collection ($n = 374$). Second, an attractive property of Bayesian statistics is their increased interpretability, allowing us to straightforwardly answer questions about the probability of our estimates to be within a certain interval. Third and finally, Bayesian methods offer a solution to the problems known as 'quasi and complete separation' (Gelman et al., 2008), which is common even with large samples sizes and few predictors (see e.g., Kimball et al., 2018), and, importantly, enables us to deal with the deterministic factor of traditional opening phrases discussed in Section 'Pictorial reference'.

A crucial aspect of Bayesian regression modelling is the specification of priors. From a linguistic perspective, there is no a priori reason to assume that mere progression of time influences the choice between indefinite or definite reference. We formalise this as a *sceptical* prior, which is centred at zero and thus assumes it is most likely that time does not have an effect. To leave some probability for possible outlier values, our sceptical prior takes the form of a Normal distribution with zero mean and scale 2.5 (cf. Ghosh et al., 2018). We employ the same conservative prior for the remaining fixed effects (*Pictorial presence* and *Opening phrase*). A similar zero mean prior is used for the Intercept with zero mean and scale 10. These priors are reasonable after scaling all regression predictors. Following Gelman et al. (2008), we centre all binary predictors to have mean 0 and centre the time predictor to have mean 0 and scale 0.5, for which we employed the R package *standardise* of Eager (2017). An additional benefit of rescaling the regression variables is that the regression output is easier to interpret.

Unlike in frequentist models, Bayesian models do not report a single, most likely coefficient β but a *distribution* of β values, called the posterior distribution θ. In our analyses, we report on the mean π, standard deviation σ, and lower and upper 95% credible interval (CI) of this distribution, which specifies the range of values in which the estimated coefficient falls with .95 probability. All regression models were executed with four chains, having 2000 iterations per chain of which the first 1000 are warm-up iterations, for each of the $n = 100$ 'overimputed' datasets, and pooled afterwards. By evaluating the mixing of the trace plots

Table 4 Summary of the results for the regression model of the wolf

	Estimate	Est. Error	l-95% CI	u-95% CI	$\hat{R}$
Intercept	0.16	0.15	−0.13	0.45	1.0
Year	0.79	0.22	0.36	1.23	1.0
Picture	0.21	0.29	−0.37	0.79	1.0

Table 5 Summary of the results for the regression model of red riding hood

	Estimate	Est. Error	l-95% CI	u-95% CI	$\hat{R}$
Intercept	−2.90	0.56	−4.16	−1.95	1.0
Year	0.26	0.31	−0.36	0.88	1.0
Picture	0.32	0.46	−0.59	1.23	1.0
Opening	−6.19	1.06	−8.60	−4.45	1.0

(which should not expose clear outliers) as well as the $\hat{R}$ statistic (which should be below 1.1, cf. Gelman and Rubin, 1992), we assessed the convergence of the models.

As a further means of evaluation, we employed projective variable selection as described by Piironen and Vehtari (2017) and implemented in the R package *projpred* (Piironen et al., 2018). Simply put, projective variable selection is a two-stage method for finding sparse, minimal subsets of features that perform well on predicting new, unseen data. The first stage revolves around finding a good performing model, irrespective of model sparsity. The second stage, then, is to find sparser models that provide similar predictions as the reference model. Validation of the variable selection process happens through approximate leave-one-out (LOO) cross-validation (see Piironen et al., 2018 for more details).

Results

If our hypothesis holds (i.e., increased cultural entrenchment of "Little Red Riding Hood" over time is reflected in increased use of definite first mention at the expense of indefinite first mention), we may expect that the odds of introducing the characters by means of definite reference increases as a function of time. As an additional hypothesis, we suggested that (i) the presence of pictures of the wolf or Red Riding Hood has a positive effect on the probability of definite first mentions, and (ii) that in the case of Red Riding Hood, the effect of time is mitigated by the co-occurrence of her introduction and traditional fairy tale opening phrases.

Table 4 presents a summary of the regression model results for story introductions of the wolf. The mean estimate of "year" equals an odds ratio of 2.2 (95% credible interval, CI [1.4, 3.4]). The proportion of posterior samples smaller than zero ($\theta < 0$) equals 0, which indicates that the probability of a negative effect is virtually absent. As such, there is ample evidence that the estimated probability for definite introductions of the wolf increases steadily over time. The mean estimate of "picture" equals an odds ratio of 1.2, suggesting a small positive effect on definite first mentions of the wolf. Note, however, that the ratio falls in the 95% credible interval of [0.7, 2.2] indicating that there is substantial probability mass against the hypothesis of a positive effect. The probability of a negative estimate for "picture" is approximately 0.24, and, inversely 0.76 for a positive effect (corresponding to an evidence ratio of 3.23),[7] which suggests mild evidence of a positive effect. As such, both estimates seem to positively contribute to the probability of definite first mentions of the wolf. However, a post-hoc projective variable selection procedure indicates that the presence of pictures does not significantly alter the model's predictive performance.

Table 6 Summary of the results for the regression model of opening phrases

	Estimate	Est. Error	l-95% CI	u-95% CI	$\hat{R}$
Intercept	0.40	0.11	0.20	0.61	1.0
Year	0.30	0.21	−0.11	0.72	1.0

How do these results compare to the first mentions of Red Riding Hood? The results are presented in Table 5. The predictor year has a mean estimate of 0.26 which corresponds to an odds ratio of 1.3 in favour of definite introductions of the little girl. Note, however, that the estimate falls in the 95% odds credible interval of [0.7, 2.4], which suggests the absence of an effect, or at least considerable probability for a negative effect. Still, the posterior probability of a positive effect ($\theta > 0$) is approximately 0.8 (evidence ratio: 3.91), which provides mild evidence for a positive effect. The presence of pictures of Red Riding Hood has a similar effect as with the wolf, with a mean estimate of 0.32 (95% CI [−0.59, 1.23]), and a 0.75 posterior probability that the estimate is positive (evidence ratio 3.03). As expected, the presence of traditional opening phrases acts as a highly strong predictor, with a mean estimate of −6.19 (95% CI [−8.6, −4.45]). To find out whether sparser versions exists of our model, we perform a projective variable selection procedure. The procedure selects opening as the sole predictor.

It appears that the effect of time plays a considerably smaller role in definite first mentions of Red Riding Hood than in those of the wolf, which is largely to be attributed to her co-occurrence with traditional opening phrases. First mentions of the wolf are not affected by these opening phrases, as the wolf is predominantly introduced mid-story when the constraints of conventional opening phrases no longer apply. Moreover, traditional opening phrases have become increasingly frequent over time, thus decreasing the probability of definite first mentions even further. The model results in Table 6 provide some statistical evidence for this time-dependent increase of conventional opening phrases. The mean estimate of 0.3 for the predictor year corresponds to an odds ratio of 1.4 (95% CI [0.9, 2.0]) in favour of opening a story with a conventional phrase, and a posterior probability of $\theta > 0 = 0.93$ (evidence ratio: 12.42) in favour of the hypothesis of a positive effect of time.

Discussion

From the analysis in Section 'Results', it appears that, controlling for pictorial presence and taking into account overruling effects such as fixed opening phrases, there is a marked increase over time in the use of definite reference to introduce the main characters in "Little Red Riding Hood". This, we argued earlier in this paper, reflects the entrenchment of the story in the sense that it reflects the author's and their audience's expectations regarding the appearance of certain characters in the story frame. In this final discussion we wish to offer some additional explanations on why we interpret these findings as such, as well as some additional reflections on what it means to observe this trend.

First, our corpus contains a collection of Dutch literary retellings of Little Red Riding Hood, which were based on the literary retellings of, among others, Perrault and the Brothers Grimm. These retellings are, in their turn, likely rooted in a non-literary, oral tradition (Tehrani et al., 2016). Thus, one could argue that a limitation of our study is that, even before the first literary retelling, oral versions of the story were already known. It is, of course, important to consider the role of oral versions when

we investigate the hypothesis that the use of definite reference reflects the cultural entrenchment of the story. Unfortunately, there is no information available concerning a possible oral tradition in the Dutch language area. However, even if we assume that such an oral tradition existed and pre-dated the literary version, the interpretation of the results remains the same. First, in the corpus used in this study, all stories can ultimately be traced back to either Perrault or the Brothers Grimm (Karsdorp and Van den Bosch, 2016), who, tellingly, both opted for indefinite reference (indicating an assessment of unfamiliarity by the authors despite any possible oral tradition). Retellings in the corpus where definite reference is chosen, then, should still be considered mutations of the base versions. Second, while our results indicate a pronounced increase of definite introduction over time, there are already cases of definite introduction even in the oldest story versions in our corpus. This may indicate that the story was already relatively well-known (through oral versions) in the Netherlands in the early 19th century. In sum, it is indeed possible that the story was already familiar to the target audience through oral version, but it remains important that there is a pronounced change in the way *literary* authors present the story frame and its characters.

Second, we wish to suggest that, if nothing else, it can straightforwardly be claimed that the observed significant increase of 'definite introduction' of the story's characters represents a change in the way which authors present (and portray) them. However, the stronger claim that the observed development is a linguistic reflection of increased cultural entrenchment is perhaps less easily digestible, depending on how much 'deliberateness' behind the linguistic choices of the author one wishes to assume. At its minimum, we should consider the level of deliberateness on the author's end in the choices they make between definite and indefinite introduction of characters at the same level as 'regular' native speakers assess and choose between definite and indefinite reference when they introduce discourse participants. At the maximum level of deliberateness, we should consider any folk and fairy tale author as a special kind of language user who deliberately manipulates the linguistic mechanism of (in)definite referent introduction to profile the story in a certain way to their audience. It is only when we assume a more or less minimal level of deliberateness that we can truly speak of the observed statistical trend as the linguistic encoding of the story's cultural entrenchment: if the author simply chooses between definite and indefinite reference at first mention of the characters based on their own assumptions of their addressee's shared knowledge (exactly as they would in everyday conversation), increased use of definite first mention reflects the accessibility of the story frame and its characters in the collective memory of author and audience. A weaker interpretation of our findings is more appropriate when we assume that the observed trend is the reflection of an emerging narrative strategy in which folk and fairy tale authors *deliberately* opt for definite introduction of characters to portray the story *as if* it is part of collective cultural memory. However, the latter, weaker interpretation only works if we assume that cultural entrenchment and collective cultural memory exist to such an extent that skilled and lesser skilled authors can wilfully manipulate their audience with it.

While it is certainly difficult to speculate on the deliberateness of the author's narrative choices, there are some indications that point towards validity of the strong interpretation. The weak interpretation would entail that there is no 'real' entrenchment, as authors merely manipulate the linguistic device of definiteness in order to signal to their audience that the characters and the story frame in which they figure are widely-known and thus important concepts, which adds value or prestige to their story. However, such manipulation could only work within a very specific context where an author presents one particular, new story to an audience. As such, we would not expect definite first mentions of folktale characters to be possible in other contexts (see e.g., Allen Ahlberg's use of definite reference in example (4)). Similarly, the claim that the use of definite reference merely serves as a narrative strategy is difficult to reconcile with the existence of fairy tale medleys (Section 'Indefinite and definite first mention of characters in narration'; Beckett, 2002, Ch. 7) and the observation that certain fairy tale characters, in particular the wolf, appear to have a conceptual status that transcends the story in which they figure (Beckett, 2008, p. 113).

Turning to the discussion of the observed developments from a Cultural Evolution stance, we should first address the question whether the increased use of definite introduction can be explained by random mutation and neutral transmission. First and foremost, it should be noted that it is very unlikely that the initial use of definite first mention for Red Riding Hood and the wolf is due to random mutation, given the minimum level of deliberate choice-making we can assume on the author's end. The remaining alternative explanation of neutral transmission, then, pertains to the subsequent spread of the feature. These so-called 'neutral models' of cultural evolution have received considerable attention in recent years as they are quite successful in explaining cultural change (see, for example, Shennan and Wilkinson, 2001; Hahn and Bentley, 2003; Bentley et al., 2007; Bentley, 2008; Ghirlanda et al., 2014; Ruck et al., 2017). In a neutral model, individuals randomly select a cultural trait for adoption, where the chance of being selected is proportional to the frequency of a trait in the population. The model thus predicts that more frequent traits are more likely to get adopted than less frequent ones. Given previous research arguing that retellings of "Little Red Riding Hood" might have been influenced by such a frequency-effect (cf. Karsdorp and Van den Bosch, 2016), an important question is whether the observed increase in definite reference can and perhaps should be attributed to such random copying behaviour, thus unintentionally favouring popular prior retellings containing definite first mentions. Indeed, in the present case, it is very difficult to ascertain whether the observed increase in definite reference should be ascribed to individual copying behaviour or to the global effect of cultural entrenchment, or, at least, there is no meaningful way to disentangle the two explanatory scenarios as they are intertwined. Constrained by a grammar's limited set of options to introduce characters to a story, new and innovative introduction strategies are not possible. As such, even if an author makes a spontaneous choice for one of the strategies (instead of copying), we cannot distinguish it from copying acts, as the number of possible introduction strategies is fixed. Thus, we cannot completely rule out the possible influence of random copying behaviour on the propagation of definite first mentions. Nevertheless, we can again counteract this interpretation by assuming that it is most likely that, in producing a new version of "Little Red Riding Hood", the author still draws on their own (native-speaker) judgements about whether or not a particular type of linguistic encoding of 'shared knowledge' is in line with their own pragmatic intuitions regarding whether the choice for that particular linguistic strategy is appropriate.

For further reflection on the subtle structural mutations observed in the lineage of "Little Red Riding Hood", it is interesting to conceptualise its reproduction history as a transmission chain in which authors continuously modify and adapt prior retellings. Artificial transmission chains have been studied extensively in Cultural Evolution, with results suggesting (i) mnemonic advantages of and social/cognitive biases for certain story elements, such as a 'negativity bias' (Bebbington et al., 2017), a stereotype consistency bias (Kashima, 2000), and (ii) selection biases for certain story types, such as a bias for

emotional stories (Heath et al., 2001) and a bias for stories describing social and survival information (Stubbersfield et al., 2015; 2017). Resulting from concentrating on simulations of oral transmission chains, an important difference between these experiments and our case is their focus on individual biases and memory limitations and how these affect the way stories are passed on from one participant generation to the next. Such memory effects are obviously not at play when stories are transmitted in written form, as in the literary versions of "Little Red Riding Hood". Furthermore, these experiments are designed in such a way that it is unlikely that participants already know the story. As such, an interesting line of future experimental research would be to investigate the effect of assumed familiarity and, more generally, assumed shared knowledge, on story transmission.

A recent study by Stubbersfield et al. (2018) more closely resembles the historical transmission chain of "Little Red Riding Hood". Using a chain design in which participants are requested to alter a written news item, thus eliminating inhibiting effects of recall, Stubbersfield et al. (2018) investigate the cultural evolution process of 'guided variation'. Guided variation is defined in Cultural Evolution as one of the main mechanisms responsible for cultural variation. Guided variation is 'directed', in the sense that it involves non-random, (un)intentional modifications and adaptations of cultural artefacts working as a "fitness-enhancing" strategy (Boyd and Richerson, 1985; Mesoudi, 2011, p. 56). It could be argued that the increase of definite first mentions as a function of cultural entrenchment can be conceptualised as an example of guided variation by pragmatic principles. A principle such as Grice's (1975) Maxim of Quantity, for instance, predicts that authors attempt to make their introductions as informative as required (and not *more* informative than required). Introducing characters by means of definite reference is "as informative as required" under the assumption that the story and its characters are accessible to the audience (from collective cultural memory). As such, introducing characters by means of definite reference can be seen as a fitness-enhancing cultural mutation, to, for example, bring the story closer to the target audience. Note that, conversely, experimental research could also reveal that the violation of conversational maxims may lead to 'unfitness' of definite introduction of folk tales: if audiences experience the use of definite first mention as an irreconcilable maxim violation on a large scale, it may be unlikely that story versions with definite introduction will win out over indefinite story versions (which use the neutral, default strategy to introduce characters in a story).

In sum, being the first to quantitatively describe the diachronic development of a story (and the concepts associated with them) into the realm of 'shared knowledge', this study has shown that it is a relatively slow and gradual (rather than immediate and abrupt) process. It is, in our view, certainly worth pursuing the multidisciplinary angle presented in this study, where insights from theoretical linguistics—in particular conversational principles of speaker-addressee dynamics as formulated in pragmatics—are put forward as explanatory factors in cultural evolution and guided variation of cultural artefacts constructed with language. For instance, besides documenting changes in cultural entrenchment (as visible in linguistic encoding), further research should also consider the possible effects of these linguistic choices on the evolution of the concepts they are meant to capture. In particular, it could be the case that the increased use of definite reference to indicate 'familiarity' with the wolf in "Little Red Riding Hood" has enabled audiences and authors to draw connections between other instantiations of 'the wolf' in other fairy tales (cf. Tehrani, 2013). The enabling of such connections could be the reason why, as noted by literary scholars, there do not appear to be several wolves, but only "one incarnation of the mythical figure of the notorious Big Bad Wolf of the fairy-tale world" (Beckett, 2008, p. 113). From an evolutionary perspective, this 'unified wolf' could be understood as a case of blending of multiple sources (i.e., from multiple story types; Uther, 2004). Conversely, the study of story transmission chains can also yield valuable insights into the linguistic study of how language users encode shared knowledge of concepts *between* versions of the same text, that is, in a 'functionally stable' reproduction lineage. Thus, there is also great potential in further extending the conceptualisation of "shared knowledge" as a dynamic entity to other text types (for instance, lineages of developing news stories) in order to further chart the (linguistic encoding of) temporality (or persistence) of different types of concepts in cultural or 'communal' memory.

Data availability

The dataset analysed in the current study is available in an online repository: https://github.com/fbkarsdorp/big-bad-wolf. This dataset was derived from the story collection repository available at: https://doi.org/10.5281/zenodo.59251.

Received: 18 November 2018 Accepted: 12 February 2019
Published online: 26 February 2019

Notes

1 A more fine-grained overview of how different linguistic forms are conventionally used to signal differences in (how speakers/writers perceive) the cognitive status of a referent is presented in Gundel et al. (1993). As the discussion in this paper focuses on 'first-mention' contexts, its scope will be limited to the distinction between proper names and definite NPs versus indefinite NPs.

2 The discussion here focuses on how first-mention definites are caused by extra-textual (or exophoric) rather than intra-textual (endophoric) reference. Exophoric reference includes reference to shared world knowledge (as discussed in this section) as well as immediate physical context (e.g., *Can you close the door (you just came through)?*). We will elaborate on exophora to physical context in Section 'Pictorial reference' when we discuss the presence of pictures in the corpus. Further note that not all definite first-mentions can be explained by means of exophora, as so-called 'unfamiliar definites' can also be triggered by intra-textual elements that require a definite article or discourse prominence (see comment in Section 'Dependent variable: (in)definite reference'). The concept of 'unfamiliar definites' is deemed irrelevant in the present study, but for in-depth discussions of the phenomenon, see, again among many others, Hawkins (1978), Prince (1992) and Epstein (2002).

3 The collection can be downloaded from https://doi.org/10.5281/zenodo.59251 (Karsdorp, 2016).

4 It should be noted, however, that the addition of pictures to a story is not necessarily something controlled by authors, but could also be the result of choices made by the editor or publisher.

5 When the bibliographic records provide a date range as estimation, we take the mean of the range as reference point.

6 We used version 2.15.0 of *brms*. All code and data can be accessed at https://github.com/fbkarsdorp/big-bad-wolf (results were obtained with the code in commit 27a69b090022072b26e08ef78b230ee2f816c438).

7 The evidence ratio is the ratio of the posterior probability of a hypothesis against that of its alternative.

References

Abbott B (2010) Reference. Oxford University Press, Oxford
Acerbi A, Kendal J, Tehrani JJ (2017) Cultural complexity and demography: the case of folktales. Evol Human Behav 38(4):474–480
Bamman D, Carney M, Gillick J, Hennesy C, Sridhar V (2017) Estimating the date of first publication in a large-scale digital library. In: Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries. IEEE Press
Bebbington K, MacLeod C, Ellison M, Fay N (2017) The sky is falling: evidence of a negativity bias in the social transmission of information. Evol Human Behav 38(1):92–101
Beckett S (2002) Recycling red riding hood. Routledge, New York, London
Beckett S (2008) Red riding hood for all ages: a fairy-tale icon in cross-cultural contexts. Wayne State University Press, Detroit

Bentley RA, Lipo C, Herzog H, Hahn M (2007) Regular rates of popular culture change reflect random copying. Evol Human Behav 28(3):151–158
Bentley AR (2008) Random drift versus selection in academic vocabulary: an evolutionary analysis of published keywords. PLoS One 3(8):e3057. https://doi.org/10.1371/journal.pone.0003057
Blackwell M, Honaker J, King G (2017) A unified approach to measurement error and missing data: overview and applications. Sociol Methods Res 46 (3):303–341
Bortolini E, Pagani L, Crema ER, Sarno S, Barbieri C, Boattini A, Sazzini M, Da Silva SG, Martini G, Metspalu M, Pettener D, Luiselli D, Tehrani JJ (2017) Inferring patterns of folktale diffusion using genomic data. Proc Natl Acad Sci 114(34):9140–9145
Boyd R, Richerson PJ (1985) Culture and the evolutionary process. The University of Chicago Press, Chicago, London
Broekhuis H, Keizer E (2012) Syntax of Dutch: nouns and noun phrases, Vol. 1. Amsterdam University Press, Amsterdam
Bürkner PC (2017) brms: an R package for Bayesian multilevel models using stan. J Stat Softw 80(1):1–28
Clark HH, Marshall CR (1981) Definite reference and mutual knowledge. In: Joshi AK, Webber BL, Sag IA (eds) Elements of discourse understanding. Cambridge University Press, New York
Christophersen P (1939) The articles: a study of their theory and use in english. Munksgaard, Copenhagen
Da Silva SG, Tehrani JJ (2016) Comparative phylogenetic analyses uncover the ancient roots of Indo-European folktales. R Soc Open Sci 3(1):150645. https://doi.org/10.1098/rsos.150645
Eager C (2017) standardize: Tools for Standardizing Variables for Regression in R. R package version 0.2.1. https://CRAN.R-project.org/package=standardize. Accessed 3 Oct 2018
Epstein R (1999) Roles, frames and definiteness. In: Van Hoek K, Kibrik AA, Noordman L (eds) Discourse studies in cognitive linguistics. John Benjamins, Amsterdam, p 53–74
Epstein R (2002) The definite article, accessibility, and the construction of discourse referents. Cogn Linguist 12(4):333–378
Garcia-Fernandez A, Ligozat AL, Dinarelli M, Bernhard D (2011) When was it written? Automatically determining publication dates. In: International symposium on string processing and information retrieval. Springer, Berlin, Heidelberg, pp. 221–236
Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. Stat Sci 7:457–472
Gelman A, Jakulin A, Grazia Pittau M, Su YS (2008) A weakly informative default prior distribution for logistic and other regression models. Ann Appl Stat 2 (4):1360–1383
Ghosh J, Li Y, Mitra R (2018) On the use of cauchy prior distributions for bayesian logistic regression. Bayesian Anal 13(2):359–383
Ghirlanda S, Acerbi A, Herzog H (2014) Dog movie stars and dog breed popularity: a case study in media influence on choice. PLoS One 9(9):e106565–5. https://doi.org/10.1371/journal.pone.0106565
Givón T (1984) Syntax: a functional-typological introduction, vol. 1. John Benjamins, Amsterdam
Grice HP (1975) Logic and conversation. In: Cole P, Morgan JL (eds) Syntax and semantics: 3. Speech acts. Academic Press, New York, p 41–58
Gundel JK, Hedberg N, Zacharski R (1993) Cognitive status and the form of referring expressions in discourse. Language 69(2):274–307
Hahn M, Bentley RA (2003) Drift as a mechanism for cultural change: an example from baby names. Proc R Soc B: Biol Sci 270:S120–S123
Hawkins JA (1978) Definiteness and indefiniteness. Croom Helm, London
Hawkins JA (1991) On (in)definite articles: implicatures and (un)grammaticality prediction. J Linguist 27(2):405–442
Heath C, Bell C, Sternberg E (2001) Emotional selection in memes: the case of urban legends. J Personal Social Psychol 81:1028–1041
Hickmann M, Kail M, Roland F (1995) Cohesive anaphoric relations in French children's narratives as a function of mutual knowledge. First Lang 15 (45):277–300
Hoffman MD, Gelman A (2014) The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. J Mach Learn Res 15(1):1593–1623
Horn LR, Abbott B (2010) <the, a>: (in)definiteness and implicature. In: Campbell JK, Kabasenche W, O'Rourke M (eds) Reference and referring, topics in contemporary philosophy, Vol. 10. MIT Press, Cambridge
Hutchinson S, Louwerse M (2018) Extracting social networks from language statistics. Discourse Process 55(7):607–618
De Jong F, Rode H, Hiemstra D (2005) Temporal language models for the disclosure of historical text. In Humanities, computers and cultural heritage: Proceedings of the XVIth International Conference of the Association for History and Computing (AHC 2005)
Kail M, Hickmann M (1992) French children's ability to introduce referents in narratives as a function of mutual knowledge. First Lang 12(34):73–94
Kashima Y (2000) Maintaining cultural stereotypes in the serial reproduction of narratives. Personal Social Psychol Bull 26(5):594–604
Karsdorp F (2016) Story-network-data: Final release (Versionv1.1). Zenodo. https://doi.org/10.5281/zenodo.59251.
Karsdorp F, Van den Bosch A (2016) The structure and evolution of story networks. R Soc Open Sci 3(6):160071. https://doi.org/10.1098/rsos.160071
Kimball A, Kailen Shantz E, Eager C, Roy J (2018) Confronting quasi-separation in logistic mixed effects for linguistic data: a Bayesian approach. J Quant Linguistics https://doi.org/10.1080/09296174.2018.1499457
Kulkarni V, Tian Y, Dandiwala P, Skiena S (2018) Simple neologism based domain independent models to predict year of authorship. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 202–212
Kumar A, Lease M, Baldridge J (2011) Supervised language modeling for temporal resolution of texts. In: Proceedings of the 20th ACM international conference on Information and knowledge management. ACM: Glasgow, Scotland, UK, pp. 2069–2072
Louwerse MM, Zwaan RA (2009) Language encodes geographical information. Cogn Sci 33:51–73
Mesoudi A (2011) Cultural evolution. How Darwinian theory can explain human culture and synthesize the social sciences. University of Chicago Press, Chicago, London
Milsark G (1974) Existential sentences in english. MIT, Cambridge, Doctoral dissertation
Piironen J, Vehtari A (2017) Comparison of Bayesian predictive methods for model selection. Stat Comput 27:711
Piironen J, Paasiniemi M, Vehtari A (2018) Projective inference in high-dimensional problems: prediction and feature selection. arXiv preprint arXiv: 1810.02406
Poesio M, Vieira R (1998) A corpus-based investigation of definite description use. Comput Linguist 24:183–216
Prince EF (1992) The ZPG letter: subjects, definiteness, and information status. In: Mann WC, Thompson S (eds) Discourse description: diverse analyses of a fund-raising text. John Benjamins, Amsterdam/Philadelphia, p 295–325
Recchia G, Slater A, Louwerse M (2014) Predicting the good guy and the bad guy: attitudes are encoded in language statistics. In: Proceedings of the Annual Meeting of the Cognitive Science Society. pp. 1264–1269
Ross RM, Greenhill SJ, Atkinson QD (2013) Population structure and cultural geography of a folktale in Europe. Proc R Soc London B 280(1756):20123065. https://doi.org/10.1098/rspb.2012.3065
Ross RM, Atkinson QD (2016) Folktale transmission in the Arctic provides evidence for high bandwidth social learning among hunter-gatherer groups. Evol Human Behav 37(1):47–53
Ruck D, Bentley RA, Acerbi A, Garnett P, Hruschka DJ (2017) Role of neutral evolution in word turnover during centuries of english word popularity. Adv Complex Syst 20(06n07):1750012. https://doi.org/10.1142/S0219525917500126
Russell B (1905) On denoting. Mind, 14:479-493. Reprinted in Marsh R C (ed) Logic and Knowledge. George Allen and Unwin, London
Sapir E (1921) Language: An introduction to the study of speech. Hartcourt, Brace and company, New York
Schmerse D, Lieven E, Tomasello M (2015) Young children use shared experience to interpret definite reference. J Child Lang 42(5):1146–1157
Schneider P, Dubé R (1997) Effect of pictorial versus oral story presentation on children's use of referring expressions in retell. First Lang 17(51):283–302
Shennan S, Wilkinson J (2001) Ceramic style change and neutral evolution: a case study from Neolithic Europe. Am Antiq 66(4):577–593
Stubbersfield J, Tehrani JJ (2013) Expect the unexpected? Testing for minimally counterintuitive (MCI) bias in the transmission of contemporary legends: a computational phylogenetic approach. Social Sci Comput Rev 31(1):90–102
Stubbersfield JM, Tehrani JJ, Flynn EG (2015) Serial killers, spiders and cybersex: social and survival information bias in the transmission of urban legends. Br J Psychol 106(2):288–307
Stubbersfield JM, Tehrani JJ, Flynn EG (2017) Chicken tumours and a fishy revenge: evidence for emotional content bias in the cumulative recall of urban legends. J Cogn Cult 17(1-2):12–26
Stubbersfield JM, Tehrani JJ, Flynn EG (2018) Faking the news: intentional guided variation reflects cognitive biases in transmission chains without recall. Cult Sci J 10(1):54–65. https://doi.org/10.5334/csci.109
Tehrani JJ (2013) The phylogeny of little red riding hood. PloS One 8(11):e78871. https://doi.org/10.1371/journal.pone.0078871. online publication 13 November
Tehrani JJ, Nguyen Q, Roos T (2016) Oral fairy tale or literary fake? Investigating the origins of Little Red Riding Hood using phylogenetic network analysis. Digit Scholarsh Humanit 31(3):611–636
Uther H-J (2004) The types of international folktales. A classification and bibliography. Parts I–III. Folklore Fellows Communications, Helsinki
Zipes J (1993) The trials and tribulations of little red riding hood, 2nd edn. Routledge, New York, London

Acknowledgements
The authors thank Marten van der Meulen for his contribution to collecting the data collection.

Additional information
Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://www.nature.com/reprints

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