Multilingual Generation of Uncertain Temporal Expressions from Data: A Study of a Possibilistic Formalism and its Consistency with Human Subjective Evaluations

Albert Gatt[a,*], François Portet[b,c]

[a] *Institute of Linguistics, University of Malta.*
[b] *Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France*
[c] *CNRS, LIG, F-38000 Grenoble, France*

Abstract

In NLG systems, temporal uncertainty in raw data can hamper the inference of temporal and causal relationships between events and thus impact the quality of the generated texts. In this paper, we introduce a framework to represent and reason with temporal uncertainty based on possibility theory and propose a model that uses the outcomes of such temporal reasoning to select linguistic expressions to convey uncertainty to the reader. Our model is based on Fuzzy Temporal Constraint Networks (FTCN) and our work is based on the assumption that uncertainty should be communicated to an end user. The model we propose is grounded in experimental data from three languages. We present a large-scale empirical study that investigates the conditions that influence human subjective uncertainty in reasoning about temporal relations. Based on this, we also construct a classifier to select expressions to convey uncertainty, based on possibility and necessity values. We then present an evaluation which shows that the predictions of the FTCN model correlate well with human subjective uncertainty in different scenarios. An evaluation of our temporal expressions classifier also suggests good results, compared to human selection of linguistic expressions, as compared to baseline models.

Keywords:
Natural Language Generation, Temporal Uncertainty, Epistemic Modality, Possibility theory

1. Introduction

Nowadays, all big collections of data are stored in distributed databases. The data may be related to business transactions, sensor data, administration, health care, etc. These data are often riddled with uncertainty or imprecision for different reasons. For instance, a patient database may contain records of interventions which were entered well after they actually occurred [1].

This problem is particularly acute in systems where the temporal dimension of the data is important and is exacerbated by the lack of a principled way of handling temporal information in

*Corresponding author. Tel.: (+356) 2340 2150
Email addresses: albert.gatt@um.edu.mt (Albert Gatt), francois.portet@imag.fr (François Portet)

existing database management systems [2]. Temporal uncertainty – that is, uncertainty about the precise time at which an event occurred – can affect the quality of data processing and inference since it has an impact on temporal and/or causal relationships between events. When decision support systems (DSS) must exploit such data, they either restrict output to what the system is most certain about, or communicate findings using a ranking mechanism [3]. Other DSSs rely on visualisation techniques to present a global view to the user. In contrast to such systems, data-to-text approaches seek to present data to users in the form of natural language summaries. Such systems use Natural Language Generation (NLG) techniques [4, 5] to describe salient aspects of data in languages such as English, following a number of processing stages whereby significant events are detected and interpreted in the raw data [6]. Such systems have been used successfully in a variety of domains, ranging from the generation of weather and meteorological reports [7, 8, 9], to medical data [10, 11, 12], as well as other types of raw sensor data [13, 14].

The present work falls within this general area and stems from work on the the family of BabyTalk systems [15, 1, 16], which provided descriptive textual summaries of heterogeneous data (entered both automatically and manually) related to individual patients in a Neonatal Intensive Care Unit (NICU), with the aim of drawing attention to relevant medical events while leaving it up to the medical professional to interpret them and make appropriate clinical decisions.

In BabyTalk and other data-to-text systems, robustness and effectiveness partly depend on the extent to which they incorporate a principled approach to temporal representation and uncertainty. To be an effective decision support tool, a summary must permit the reader to reconstruct the temporal sequence of the events that it narrates, and make clear the relations between them. Furthermore, where the precise time at which events occurred is not available, the inference of temporal and/or causal relationships between events can be compromised [17].

In this paper, we argue that temporal uncertainty in data, or uncertainty resulting from inferences, should be explicitly communicated in textual outputs. Failure to do this can result in erroneous decision-making, whose consequences can be serious in some application areas. Natural language provides a number of mechanisms for communicating such uncertainty. On the other hand, it is by no means clear how a formalism to reason with uncertainty would map to such linguistic expressions. Furthermore, it is not clear whether the same mapping would work for different languages, given that languages display a variety of such expressions.

This paper addresses two empirical questions which are of direct relevance to the expression of temporal uncertainty to text generation. The first concerns the representation and quantification of uncertainty: given the raw data about an event, as well as general knowledge that enables a limited amount of reasoning about a situation, we are interested in quantifying the degree of uncertainty about the time of an event and the resulting degree of uncertainty about the temporal relations between it and other events (e.g *x happened before y*). We propose to use a soft computing approach to handle this, showing that its predictions have a good correspondence to human intuitions.

Our second question focusses on the use of modalised propositions[1] and on the way modal expressions can be *grounded* in subjective uncertainty arising from raw data. We describe an experimental design, replicating and extending an experiment reported in [19], that enables us both to quantify subjective uncertainty in a given situation, and to map from subjective uncertainty to modal expressions. Our experiments are conducted in three different languages which, though culturally fairly close (insofar as they are European), are typologically diverse. In this way, we

[1] In what follows, our use of the term 'modality' refers to the semantic or 'notional' category [18].

seek both to validate our methodology using data from multiple languages, and to investigate the implications that differences between languages can have for a proper account of modality in NLG.

We begin with an overview of related work (Section 2), followed by a description of the reasoning formalism (Section 3). We then describe an experiment that elicited human judgements of temporal uncertainty as well as corresponding linguistic expressions (Section 4). In Section 5, we re-use the experimental methodology to evaluate the model proposed in Section 3. We conclude in Section 6 with some pointers to future work.

2. Epistemic uncertainty in language

Consider the following hypothetical scenario. Suppose that an NLG system produces a summary of events extracted from raw data, and it is only possible to ascertain that event e_1 happened within a particular temporal interval $[t_1, t_2]$. Now consider the situation where the system determines that e_1 may be in some temporal relationship to e_2. Given that the time of e_1 is a uncertain temporal interval, this temporal relationship cannot itself be determined with certainty. Should the system merely assert the relationship (e.g. e_1 *happened before* e_2), it would risk misleading the reader into thinking that the relationship holds, without qualification. On the other hand, neglecting to mention the relation, due to the uncertainty surrounding it, may cause the reader to conclude that no such relationship holds. An alternative would be to communicate to the reader that, as far as the system 'knows', such a relationship is *possible* (e.g. through an expression such as e_1 may *have occurred before* e_2).

Uncertainty about the time of an event, and therefore about the possible temporal relations between it and other events, is a form of *epistemic* uncertainty, that is, its source resides in the speaker's (or NLG system's) state of knowledge. In natural language, a prominent way to express such uncertainty is through the class of expressions known as epistemic modal expressions, which are concerned with the degree of possibility or necessity associated with a particular proposition. Modality can be characterised in terms of assertion [20]. Indeed, while an unmodalised proposition is often simply asserted (thereby presupposing certainty on the part of the speaker about the matter), its modalised counterpart is not, or only with some qualification as to the degree of evidence that the speaker has for it.

From the perspective of a data-to-text NLG system, the use of epistemic modality to express temporal uncertainty gives rise to two challenges. The first is to actually quantify the degree of uncertainty, an issue we turn to in Section 3 below. The second is to choose the 'right' expression *given the uncertainty*. This presents a system with a choice problem: within a language, the class of epistemic modal expressions can be quite broad; across languages, there is also considerable diversity.

Both these challenges can be illustrated with the following example.

(1) A bank robbery occurred yesterday afternoon. An investigator is trying to reconstruct the scene from eye-witness reports. He knows for certain that the robbers were inside the bank for no more than 45 minutes. He also knows for certain that the police took 30 minutes to arrive on the scene after being alerted. He has also interviewed some eye-witnesses. Here is what they said: **The robbers entered the bank at 16:00. The police were alerted some time between 16:15 and 16:45.**

Consider now the proposition *The police were on the scene before the robbers left the bank*. In this scenario, the certainty of this proposition is affected by the fact that the event of the police

being alerted occurs within an uncertain interval. From an NLG perspective, we would like to be able to (a) quantify the degree of certainty associated with the occurrence time of the two events, as well as their temporal relation; and (b) choose the right expression to express this. A prerequisite for both these tasks is a computationally tractable account of how modal expressions are grounded in temporal data, which also supports fine-grained choices, such as that between *may* and *possibly*.

However, it is unlikely that a model of such choices can be built completely language-independently, since modality exhibits considerable cross-linguistic variation [20]. Languages like English and French would commonly modalise a proposition using modal auxiliaries (2a) or adverbials (2b). Whether the two systems (auxiliaries and adverbials) are equivalent with respect to the degree of uncertainty they express is an empirical question, one that has a direct impact on the lexicalisation strategies used by an NLG system.

(2) (a) *La police* **pourrait/doit** *avoir été sur les lieux*
the.fsg police may.3pl/must.3pl have be.3sg.ps on def.pl scene.pl
avant que les voleurs quittent la banque.
before the.pl robber.pl leave.3pl.ps the.fsg bank
'The police **may/must** have been on the scene before the robbers left the bank.' The police **may/must** have been on the scene before the robbers left the bank.

(b) *La police était* **surement/peut-être** *sur les lieux avant que*
the.fsg police be.3sg.ps definitely/possibly on def.pl scene before
les voleurs quittent la banque.
the.pl robber.pl leave.3pl.pl the.fsg bank
'(**Possibly**) the police were (**definitely**) on the scene before the robbers left the bank.'

Remark: The first and third lines of each example are the two versions of a same event in French and English. The second line is the word-for-word translation in English of the first line with grammatical hints (pl = plural, sg = singular, def = definite, 3sg = third person of the singular, etc.)

The above example suggests certain similarities between English and French, despite their different typological classification (Anglo-Saxon vs. Romance). The difficulties increase when other language families are considered. We will also consider Maltese, a European language which comes from a third language family, Semitic, where the modal auxiliaries that have been identified [21] tend to be more restricted in their use. For example, the auxiliary *seta'* (can.3sgm.pfv; 'could have') can be used to express epistemic possibility or likelihood, but so can adverbs like *bilfors* (e.g. *bilfors li*; lit. 'by force that', i.e. 'definitely') and *żgur* ('certainly'; cf. example 3a) and the Romance-derived *forsi* ('maybe/perhaps'; cf. example 3b).

(3) (a) *Il-pulizija* **jista' jkun/bilfors/żgur** *li kienu fuq ix-xena qabel ma*
the-police could be/definitely/certainly that be.pl.ps on the-scene before
l-ħallelin telqu mill-bank.
the-robber.pl leave.pl.ps from.the-bank

'The police **may have/definitely/certainly** left the scene before the robbers left the bank.'

(b) *Il-pulizija* **forsi** *kienu fuq ix-xena qabel ma l-ħallelin*
the-police possibly be.3pl.ps on the-scene before the-robber.pl
telqu mill-bank.
leave.3pl.ps from.the-bank
'**Possibly** the police were on the scene before the robbers left the bank.'

The examples from the three languages under consideration serve to illustrate a subset of the grammatically diverse expressions that different languages make available to express epistemic uncertainty, as well some possible differences that may arise among them despite their cultural proximity (insofar as all three are European languages). A consideration of languages which are even more diverse – historically, culturally and typologically – would presumably shed light on even greater differences in modal systems and their interaction with the expression of time, in line with recent work that questions the existence of absolute 'universals' across languages [22]. An investigation of such cross-linguistic differences is beyond the scope of the present paper, though the methodology illustrated in the following sections is not restricted to particular languages.

Neither of the two questions we have raised – that of representing and quantifying uncertainty, and that of mapping from this to the right modal expression in a particular language – has been treated exhaustively in the NLG literature. To our knowledge, the only recent approach to handling modals in NLG is [23], which focuses on the generation of deontic modals (those related to obligation, rather than epistemic certainty), in the CAN system, which advises students about university courses [24, 23]. Klabunde's approach is based on a possible worlds framework [18], in which the truth of a modalised proposition is evaluated against a contextually determined set of relevant possible worlds or situations, ordered by their accessibility from the current world or situation. In an epistemic context, this set contains the worlds which are compatible to some degree with the propositions which constitute the underlying 'evidence' for the statement. Thus, in the CAN system, the conversational background is determined by the domain knowledge about available courses. Given a user query about a course, the planner identifies several alternative plans that lead to the same goal – essentially an operationalisation of the notion of a 'set of possible worlds'. Modal force is computed by quantifying over plan nodes: for example, if all alternative plans leading to the taking of a given course involve a prerequisite, then the system selects the German equivalent of *must* in expressing this prerequisite to the user.

Traditionally, semantic work on modality has been based on the possible worlds framework (e.g. [18, 25]; but see [26, 27] for alternative accounts). However, a possible worlds approach is not straightforwardly applicable to the type of problem illustrated in (1). Intuitively, the temporal uncertainty of the proposition in the example, which arises due to an event having a fuzzy temporal interval, would be evaluated on a continuous scale: given the knowledge that something occurred between times t_1 and t_2, a person may feel more certain of the occurrence towards the middle of the interval, less so as one approaches its start or end. If a continuous certainty scale is what is required, it is difficult to see how approaches based on a treatment of propositions as (crisp) sets of possible worlds can be applied. In the following section, we consider an alternative proposal.

3. Temporal representation and reasoning

In AI, temporal representation and reasoning systems are usually based on actions and changes, or on temporal constraints. The former approach is devoted to reasoning with sequences of actions that change the state of the world while the later deals with relations among temporal objects (e.g., instants and intervals). The most common relations considered in the temporal constraints approach are the qualitative ones (e.g., "James was walking while it was raining") and the metric ones (e.g., "Mary left 10 minutes before James arrived"). Qualitative constraints are usually based on Allen's thirteen mutually exclusive binary relation [28]. Examples of metric-based models include Metric Temporal Constraint Networks (TCN) [29], which specify numerical temporal relations between pairs of time points and make it possible to compute the consistency of the network through a specific constraint propagation algorithm.

Our aim is to reason about temporal data to extract relevant information to be expressed linguistically. The formalism we use to represent and reason with events and relations between events is based on the TCN [29] approach which is capable of subsuming the qualitative one and represents events as time-points rather than intervals. However, the time-point metric approach is capable of representing intervals through start and end points and can translate most qualitative intervals or point relations into metric relations (e.g., a before b can be reformulated as $b - a \in [1, \infty)$), though this may reduce the expressiveness of the interval relations [30]. Moreover, there are numerous algorithms to compute the consistency of a TCN network efficiently, depending on the level of expressiveness allowed, though expressive power and computational tractability tend to be inversely related. Other interesting properties of TCNs are that they can be used to represent numerical temporal information that can then be queried or used to model expert knowledge [31, 32].

In the TCN formalism, temporal representation relies on time points and time is considered as a linearly ordered discrete set of instants ($t_0 < t_1 < \cdots < t_i < \dots$) where $\forall i \in \mathbf{N}$, $t_{i+1} - t_i = \Delta_t$. Δ_t is a constant that represents the sampling period (e.g. 1 microsecond, 1 month, 1 century). We assume that temporal information is composed of instantaneous events and finite durative events (i.e. intervals but not semi-intervals) as well as temporal constraints between events. In the following, we use lowercase italics (a,b,c,...) for instantaneous events, and normal uppercase for intervals (A,B,C,...).

Let us assume that events being described by an NLG system are classified according to some ontology O, which defines the possible types in the world under discussion. An instantaneous event or **event** a is a tuple $\langle t, o \rangle$, where $t \in \mathbf{N}$ and $o \in O$. t is the known date of occurrence of the event and o represents some structured data corresponding to this event (e.g. an instance of a type defined in O, based on a database record, inference made by the system, or user input) [2].For instance, "the robbers entered the bank at 16:00" can be represented by $\langle 16{:}00, \text{“}ROBBERS\_ENTER\_BANK\text{”} \rangle$.

A durative event or **interval** A is a tuple $\langle s, e, T, o \rangle$, where s (resp. e) is an instantaneous event representing the start (resp. end) of the durative event, T is a temporal constraint such that $e - s \in T$ and o is the description of the durative event. For instance, "the police took 30 minutes to arrive" is represented by:

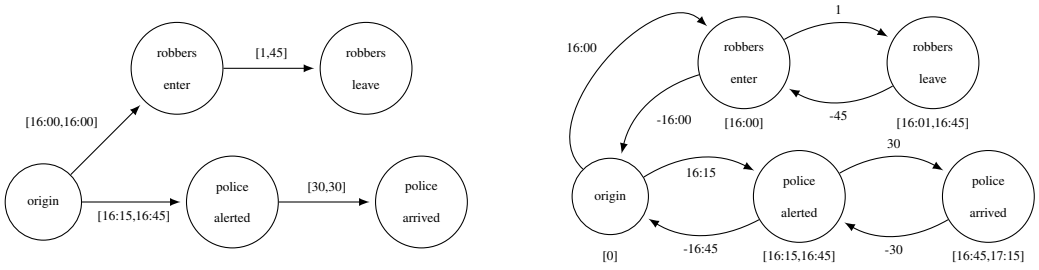
$$\langle \langle time\_of\_alert, \text{“}POLICE\_ALERTED\text{”} \rangle, \langle time\_of\_arrival, \text{“}POLICE\_ARRIVED\text{”} \rangle, [30, 30], \text{“}POLICE\_COMING\text{”} \rangle.$$

[2]The information o will not be used for temporal reasoning, but its representation is important to processing chain of textual generation.

Briefly, a TCN $\mathcal{N}$ is a graph whose nodes are instantaneous events $(a, b, c \dots)$ and edges are temporal constraints between them. Each constraint $T_{a,b}$ between event a and event b is represented by a set of disjunctive binary constraints $(\{I_1, \dots, I_n\} = \{[t_{s1}, t_{e1}], \dots, [t_{sn}, t_{en}]\})$,[3] which implies $(t_{s1} \leq b - a \leq t_{e1} \vee \dots \vee t_{sn} \leq b - a \leq t_{en})$. It is assumed that constraints are given in a canonical form where all intervals are pairwise disjoint. In our approach all durative events are represented in the TCN as two events (nodes) constrained by the temporal duration (edges).

A set of values $V = (v_a, v_b, v_c, \dots)$ is a solution to the TCN if the assignment $\{a = v_a, b = v_b, c = v_c, \dots\}$ satisfies all the constraints. A value v is a feasible value for variable a if there exists a solution in which $a = v$. The set of all feasible values of a variable is called the minimal domain. A TCN is consistent if a least one solution exists.

For instance, the set of facts in example (1), can be represented by the TCN depicted in Figure 1a where all durative events are translated into pairs of events (e.g. 'were inside the bank' → 'robbers enter' and 'robbers leave') and all temporal relations are translated into binary temporal constraints (e.g., 'for no more than 45 minutes' → $[1, 45]$). This also applies to absolute times (i.e., unary constraints such as $a <$ 16:00), which are represented with respect to the origin of the day (i.e., represented as a binary constraint $origin \leq a - origin <$ 16:00 $- origin$; for simplicity we will assume $origin = 0$) .



a) TCN

b) distance digraph with origin fixed at 0 and minimal domain of events

Figure 1: Robbers example represented as a TCN.

In the TCN approach, reasoning is seen as a temporal constraint satisfaction problem (TCSP), which consists in finding a solution that satisfies a set of inequalities (e.g., $t_{s1} \leq b - a \leq t_{e1} \vee \dots \vee t_{sn} \leq b - a \leq t_{en}$). In practice, this is achieved by converting the TCN into a weighted directed distance graph and applying algorithms that solve the shortest path problem (such as the Floyd-Warshall one) to generate the minimal network (i.e., the network with the tightest constraints). For instance, Figure 1b shows the digraph equivalent to the TCN of Figure 1a after computation of the minimal domain of each temporal event. It can be seen that the feasible values for the event 'robbers leave' are those of the interval [16:01,16:45]. In this case the network is said to be consistent, since none of the events has an empty minimal domain. For more information about temporal reasoning and the aforementioned formalisms the reader is refereed to [33, 34].

The TCN can also be used to test the validity of some hypotheses. For instance, if one wants to test the assertion *The police were on the scene before the robbers left the bank*, this constraint

[3]Formally, unary constraints are allowed but they are generally represented with a binary constraint with an origin event *ori*, e.g., $a \in [12{:}03, 12{:}45] \rightarrow 12{:}03 \leq a - ori \leq 12{:}45$ where *ori* is the beginning of the day

can be integrated into the network (before $\rightarrow [1, \infty)$; see the dashed edge in Figure 2). The addition of this constraint makes the network inconsistent since the latest possible departure time of the robbers is 16:45 and the earliest police presence is 16:45, which is not strictly before the robbers' departure.

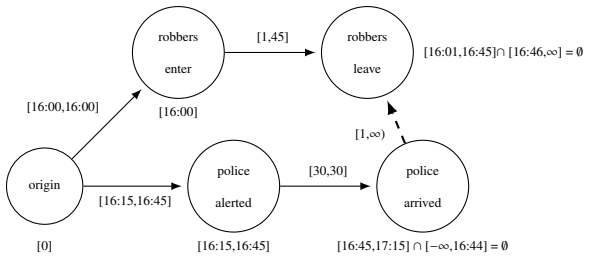


Figure 2: Robbers example represented as a TCN when the constraint 'before' [1,∞) is inserted.

While such reasoning is perfectly correct, it might not correspond to the intuitive answer a human would give. A human reader is likely to take much more liberty with the interpretation of the reported temporal facts, particularly if it is a report made by another person. For instance, the statement that the police took 30 minutes to arrive might result in some allowance being made for their arriving after 29 minutes, or after 31. A slight change in the interpretation of the constraints would lead to very different results. Temporal constraints account for uncertainty in temporal knowledge only to a certain extent. However, this is not sufficient for domains where temporal knowledge is highly pervaded with vagueness and uncertainty. To better capture these intuitions on human reasoning about time, it is possible to represent some temporal constraints as a fuzzy sets [35].

TCNs dealing with fuzzy temporal constraints are known as *Fuzzy Temporal Constraint Networks* (FTCN) [36, 37, 38]. Several implementations of such models exist, but for the sake of conciseness, we will focus on the one implemented in the FuzzyTIME engine developed by the AIKE group of the University of Murcia [36, 38]. FuzzyTIME is a general purpose engine that can represent intervals as well as instants and all common qualitative and quantitative temporal relations between them. All definitions are translated into metric relations between time points on which the reasoning is performed. FuzzyTIME does only allow convex normalized distribution. This constraint, although limiting the expressive power of the language, allows the minimization algorithm to be executed in polynomial time $O(n^3)$ [39].

In this approach, a binary constraint $T_{a,b}$ between two events a and b is defined by a normalised, unimodal possibility distribution $\pi_{T_{a,b}} \in \boldsymbol{Z}$ which restricts the possible values of the time elapsed between a and b. When no other constraints and variables are present, the assignment $a = v_a$ and $b = v_b$ is a solution only if $\pi_{T_{a,b}}(b - a) > 0$.

A FTCN is thus a TCN in which all temporal constraints are treated as fuzzy temporal constraints. A set of values $V = (v_a, v_b, v_c, \dots)$ is a $\sigma - possible$ solution if the assignment $\{a = v_a, b = v_b, c = v_c, \dots\}$ satisfies all the constraints with at least possibility σ.

Recall that in possibility theory [40], the uncertainty about a temporal relation r between two events a and b can be evaluated by the two dual measures of possibility Π and necessity (also called certainty) N, as follows:

$$\Pi(r_{a,b}) = \pi_r(b - a) \tag{4}$$

$$N(r_{a,b}) = 1 - \Pi(\bar{r}_{a,b}) \tag{5}$$

where $\pi_r(b-a) \in [0,1]$ is the possibility distribution of the temporal distance between the events a and b, representing the degree to which these two events are possibly linked via relation r, and $\bar{r}_{a,b}$ is the complement of $r_{a,b}$.

The necessity of the relation r between a and b can be summarised as follows: $r_{a,b}$ is certain only if no relation contradicting $r_{a,b}$ (i.e., $\bar{r}_{a,b}$) is possible. If several contradictory relations are completely possible at the same time (e.g. *before* and *after* are completely possible), no certainty exists.

An example FTCN is represented in Figure 3 where the arrival time of the police is translated into a possibility distribution expressing the following interpretation: *it is completely possible that the police took 30 minutes to arrive, less possible that they took 28-30 minutes or 30-32 minutes, and impossible otherwise*. In this example, all other constraints are represented as a uniform possibility distribution (e.g., the constraint $[1, 45]$ is translated into a possibility distribution for which any value in its range is completely possible, i.e., a crisp interval).

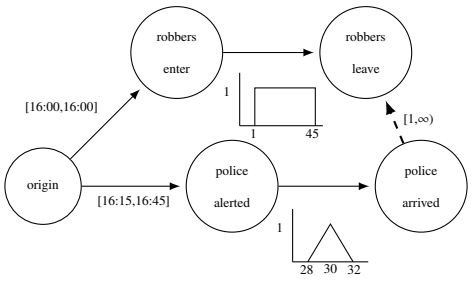


Figure 3: Robbers example represented as a FTCN.

In FTCN, the solutions to the network can satisfy the constraints only to a certain possibility degree σ, given that temporal constraints may be fuzzy. In FuzzyTIME, an algorithm that combines exhaustively all constraints is applied to obtain the minimal network (i.e., in which the constraints have the smallest possible degree of imprecision) [36]. For instance, incorporating the assertion *The police were on the scene before the robbers left the bank.* with Δ_t = 1 minute leads to a network consistent with only .5 possibility and 0 necessity (because the complement relation 'after' is completely possible).

This model therefore offers a way of quantifying the possibility and necessity of an event, given a formalisation of the background knowledge. Thus, this formalism can handle the first of the two problems pointed out in the previous section, namely, to quantify temporal uncertainty of events in a fine-grained manner. Our next question is how these values can be mapped to linguistic expressions by an NLG system. We first present the results of an empirical investigation into the relationship between the formal notions of possibility/necessity and their corresponding expressions in natural language. Subsequently, we discuss an implementation of the model presented here to deal with these mappings, followed by an evaluation.

4. Expressing temporal uncertainty: a cross-linguistic investigation

The experiment described in this section was conducted with a view to mapping subjective certainty judgements to the classes of modal expressions in French, Maltese and English

introduced in Section 2. Two main research questions were addressed. First, does a possibility-theoretic formalism, as described in the previous section, adequately capture subjective uncertainty judgements of temporal uncertainty by humans? Second, is it a viable model to use as the underlying reasoning mechanism for an NLG system, on the basis of which a modal expression can be selected during the generation of a linguistic description of data? Given the cross-linguistic diversity of such expressions, it is important to ascertain this in more than one language.

The experiment replicates and extends a methodology reported by [41, 19], albeit with substantial modifications in the choice of materials, and with the crucial difference that it was carried out on three groups of native speakers of the three languages under consideration. It was conducted online, using a 'crowdsourcing' methodology. Participants were presented with scenarios such as (1) above and were asked to rate their subjective certainty regarding a specific temporal relationship, and also to choose from a set of linguistic expressions that would express this uncertainty. Before discussing the details of the experiment, we first discuss some of the methodological challenges in conducting a study of this nature.

4.1. Eliciting uncertainty judgements: the Ψ measure

Getting judgements of uncertainty from human volunteers about uncertain temporal data is challenging in at least two respects. First, the temporal problem presented to the participant must allow him/her to understand the question and carry out temporal reasoning without incurring too much cognitive load. This is an issue we sought to address in our construction of experimental materials, described in Section 4.5 below. Second, human judgements have to be elicited using a scale that is intuitive. Since possibility theory is largely unfamiliar to naive participants, expecting direct judgements in terms of possibility and necessity values would be unreasonable. Furthermore, since the relationship between possibility and necessity is unlikely to be transparent to such participants, we wished to avoid eliciting judgements based on two scales (one each for Π and N). Rather, we used a single scale, the Ψ measure [42], which allows the estimation of both values from a single subjective judgement.

Previous experiments by Raufaste et al. [42] have shown that eliciting separate possibility and necessity ratings from humans leads to values that do not fit the assumptions of Possibility Theory. As far as human reasoning goes, something is fully possible only if it is certain to some extent. According to the authors, this might be due to considerations that are completely separate from the formal aspects of the theories, such as the desire to avoid extreme opinions.

An alternative would be to allow any participant P to choose only one of two scales between possibility and necessity where $\Pi(P) \in [0, 1)$ and $N(P) \in [0, 1]$. If the participant chose the possibility scale then $N(P) = 0$, while if the participant chose the necessity scale then $\Pi(P) = 1$. Although this would remove inconsistency in the values, this solution is still too complex for a participant who is not familiar with Possibility Theory, since it still requires knowledge of the difference between the two measures.

Ideally, participants should provide their estimates in one measure; furthermore, this should permit a direct mapping from spontaneous confidence judgements to possibility measures, avoiding the bias due to the two measures. This measure should also support the computation of separate possibility and necessity values. Raufaste et al [42], after a suggestion by Didier Dubois, proposed to use the Ψ scale. This combines both possibility and necessity into a single scale, which ranges from 'impossible' ($\Psi = 0$) to 'completely certain' ($\Psi = 1$). From this Ψ measure, the corresponding possibility (Π) and necessity (N) values can easily be reconstructed using (6) and (7) below.

$$\Pi(P) = \begin{cases} 2 \times \Psi & \text{if } \Psi \leq 0.5 \\ 1 & \text{if } \Psi > 0.5 \end{cases} \tag{6}$$

$$N(P) = \begin{cases} 0 & \text{if } \Psi \leq 0.5 \\ 2 \times \Psi - 1 & \text{if } \Psi > 0.5 \end{cases} \tag{7}$$

This measure and the formulae (6) and (7) ensure that the constraint $N(P) \neq 0$ *iff* $\Pi(P) = 1$ is respected. It is worth noting that $\Psi(P) = 0.5$ is the equivalent of total uncertainty where everything is possible ($\Pi(P) = 1$) and nothing is necessary/certain ($N(P) = 0$). The $\Psi(P)$ measure makes it possible to provide an intuitive scale to the participants (corresponding to 'impossible' at one extreme and 'completely certain' at the other, with the middle of the scale corresponding to 'don't know') and was thus used in the following experiment.

4.2. Participants

Participants were recruited opportunistically, by advertising the online interface for the experiment in the United Kingdom, France and Malta. Each participant who visited the site was given a choice of which of the three languages (English, French, Maltese) to carry out the experiment in, depending on their native language. A total of 141 participants completed the experiment, resulting in different sample sizes for the three languages (31 English, 67 French and 43 Maltese speakers). All participants were self-reported native speakers and none were informed that the experiment was dealing with possibility theory. Table 1 displays proportions of participants by gender and language.

	Female	**Male**
English	22 (71%)	9 (29%)
French	26 (39%)	41 (61%)
Maltese	27 (63%)	16 (37%)

Table 1: Participants in the experiment.

Participants came from a variety of academic backgrounds, ranging from the arts and humanities (11%), cognitive and/or social sciences (15%), languages (26%), mathematical and computing sciences (39%) and physical or natural sciences (7%). A further 2% had no self-reported university or college education.

4.3. Procedure

The experiment exposed participants to scenarios such as those in example (1) through a web interface; this is partially displayed in Figure 4. Each scenario presented some background information, and then presented two propositions about two different *key events* (shown in boldface in (1). Key events always contained either an exact or uncertain temporal expression, which could refer to the clock time of an event (e.g. *at 16:00*, *between 16:00 and 16:45*) or to its date (e.g. *in 1890*, *between 1890 and 1895*), depending on the scenario. The scenarios were designed to make it explicit that the events themselves actually happened for certain and that uncertainty was only related to their timing. After reading a scenario, participants performed two tasks:

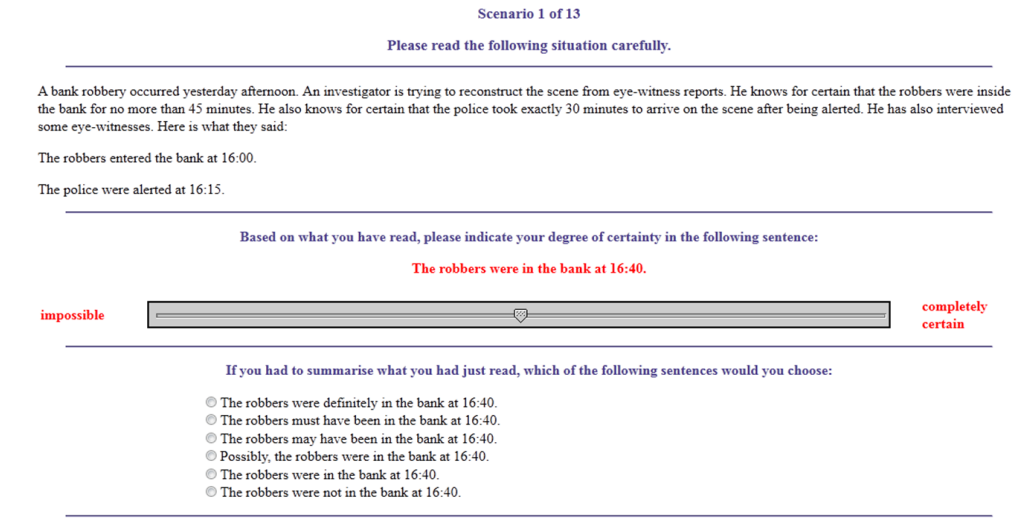


Figure 4: Partial screenshot of the experiment interface

1. *Judgement*: Participants were given a simple (i.e. *unmodalised*) proposition involving a simple event or a temporal relation between two events, and were asked to judge their subjective certainty about the proposition on a scale (Figure 4, top). To elicit these subjective certainty judgements, we used a slider representing the Ψ-scale [42] described in Section 4.1 above.
2. *Expression choice*: For each scenario, participants were also presented with a list of 6 different versions of the proposition that they had judged. These were presented in random order and participants were asked to choose the one that they felt best reflected their degree of certainty about the temporal relation expressed (Figure 4, bottom). The list invariably included the original unmodalised proposition (hereafter referred to as the *default* case), as well as a negated version. These were intended to cover the cases of complete certainty about the truth of a proposition (by hypothesis, in the conditions with no uncertainty), or about its falsity (hence, certainty that the proposition is false). Apart from these, there were 4 versions containing the expressions exemplified for the three languages in examples (2) and (3) and summarised in Table 2. Note that the expressions are grouped together in this Table based on the authors' intuitions for convenience of presentation; whether or not the expressions in the three languages correspond precisely is one of the empirical questions we sought to address.

4.4. Design

The experimental scenarios represented combinations of two within-participants factors:

- *Uncertainty* (3 levels) manipulated the amount of temporal uncertainty in a scenario. In one condition, both of the key events in the scenario were given an exact time (e.g. *at*

English	French	Maltese
must	doit	bilfors
may	pourrait	jista' jkun
possibly	peut-être	forsi
definitely	sûrement	żgur

Table 2: Modal expressions used in the experiment, in addition to the default (simple assertion) case, and negation.

16:00); in another condition, one of the key events had an uncertain time, specified as a temporal interval (e.g. *between 16:00 and 16:45*); in the third condition, both key events had such uncertain times.

- *Proposition Type* (4 levels) manipulated the type of proposition whose subjective certainty participants were asked to judge, namely: a simple proposition describing either of the two key events alone (e.g. *the robbers left the bank at 16:45*); or a compound proposition which described both key events in a temporal relation. The temporal relation was expressed using one of the following temporal connectives: *before, after*, or *during*.

This design yields $3 \times 4 = 12$ conditions. We added a thirteenth condition, in order to balance the design by ensuring that, for every level of uncertainty, there was a simple proposition describing either the first key event or the second.

In addition to the above, there was also a third, between-groups factor, namely Language (Maltese/English/French), corresponding to the language in which the experiment was conducted. The choice of language was an important part of the experiment. As argued in Section 2, expressions of uncertainty, as reflected by epistemic modals, differ across languages. In line with the research questions outlined at the start of this section, our work partly seeks to address the question whether a formal model couched in the terms of Possibility Theory can yield useful insights for NLG where expressing temporal uncertainty is concerned, irrespective of linguistic differences.

In summary, our experiment had a mixed 3 (Uncertainty) ×4 (Proposition Type) ×3 (Language) design.

4.5. Materials

Thirteen scenarios were constructed. Each scenario was rendered in English, Maltese and French translation; translations were made by native speakers and were checked to ensure minimal variation in the contents of the scenarios across the three languages. A complete list of all the thirteen scenarios used is provided in the Appendix; see (1) above for an example.

Each scenario consisted of a brief narrative that provided some background. This was followed by two events, each with a duration that could be precise or fuzzy. We restricted the presentation to two events for two reasons: (1) this is the minimum required to reason about a temporal relation; (2) we wished to avoid more than two events to avoid information overload, which could have compromised results due to significant cognitive processing being required on the part of participants to determine their degree of certainty that a particular temporal relation

	after	before	during	simple
English	0.73 (0.36)	0.73 (0.37)	0.72 (0.37)	0.81 (0.29)
French	0.83 (0.31)	0.82 (0.32)	0.61 (0.41)	0.82 (0.30)
Maltese	0.82 (0.28)	0.83 (0.29)	0.68 (0.35)	0.82 (0.23)
Overall	0.78 (0.33)	0.80 (0.32)	0.66 (0.39)	0.82 (0.29)

(a) No Uncertainty

	after	before	during	simple
English	0.54 (0.26)	0.59 (0.33)	0.72 (0.28)	0.38 (0.26)
French	0.55 (0.29)	0.60 (0.33)	0.65 (0.33)	0.39 (0.23)
Maltese	0.59 (0.26)	0.59 (0.28)	0.67 (0.27)	0.48 (0.23)
Overall	0.56 (0.29)	0.59 (0.31)	0.67 (0.30)	0.41 (0.24)

(b) One uncertain proposition

	after	before	during
English	0.50 (0.29)	0.60 (0.27)	0.64 (0.32)
French	0.57 (0.30)	0.59 (0.30)	0.64 (0.32)
Maltese	0.63 (0.29)	0.60 (0.27)	0.67 (0.29)
Overall	0.57 (0.31)	0.60 (0.28)	0.65 (0.31)

(c) Two uncertain propositions

Table 3: Mean Ψ values across languages and conditions (standard deviation in parentheses)

holds. For example determining whether three events are related in a particular way would require participants to reason about each pairwise relation among the three, which is taxing given working memory limitations [43, 44].

Rather than present temporal information in an abstract fashion (e.g., "*A* started at 16:00 and lasted 10 minutes and *B* started between 15:45 and 16:05, what is your degree of certainty that *B* started while *A* was ongoing?"), we sought to use narratives to convey the information. Since Wason and Shapiro [45], it has been shown that participants perform better at reasoning tasks when these involve concrete content, compared to cases where reasoning is based on more formal presentation of information. Furthermore, familiarity with the content enables participants to reason about actions in a manner compatible with their reasoning in real life [46]. This is why the scenarios were built so as to relate to general themes easily accessible by any adult. A further advantage of using narratives is that humans have finely developed abilities to reason and make inferences from narrative text [47, 48].

For each scenario, thirteen different versions were constructed corresponding to the conditions described above. The scenarios were rotated through a Latin square to create 13 versions of the experiment in each language, where each scenario appeared in each condition exactly once across the 13 versions.

4.6. Results

Due to a coding error, one of the scenarios for the French experiment, in the conditions where the temporal relation between the two events was *after*, was excluded from analysis.

Table 3 summarises the mean Ψ ratings overall and within each language, as a function of the different levels of Proposition Type and Uncertainty factors.

At a glance, the Table shows a clear tendency for subjective certainty to decrease as scenarios introduce more temporal uncertainty; this is most evident in comparing conditions with no uncertainty to those with one or two uncertain propositions. However Proposition Type also seems to affect Ψ ratings, at least in some conditions.

We analysed the data using Linear Mixed Effects models. Below, we report both model comparisons and estimates of significance of main effects and interactions.[4] Unless otherwise noted, all models have a maximal random effects structure, incorporating random intercepts and slopes for subjects and items [51]. In case a model with a maximal random effects structure did not converge, we simplified the random effects structure by removing slopes for interactions. Fixed factors were scaled and centred to reduce collinearity.[5]

We first construct a baseline (model 0) consisting only of the intercept, together with random intercepts and slopes for subjects and items. We then assess the contribution of the fixed effects of Proposition Type and Uncertainty (a) separately, by constructing models incorporating each one (models 1 and 2) and comparing each to the baseline; and (b) jointly, by constructing a model with both fixed effect terms (model 3) and comparing it to the one with separate fixed effects. Finally (c) we compare model 3 to a maximal model incorporating the two fixed effects and their interaction (model 4). In each case, model comparison is carried out on the basis of goodness-of-fit, using the Bayesian Information Criterion (BIC) and Log-likelihood estimates.[6] To facilitate interpretation in terms of main effects and interactions, we also report the results of an ANOVA over the maximal model incorporating the fixed effects and their interaction. Finally, we also report on a model (5) that extends the maximal model by incorporating the fixed effect of Language. Table 4 summarises all models and comparisons.

The model comparisons suggest that Uncertainty plays a crucial role in affecting subjective Ψ ratings, as shown by the better fit to the data of Model 1, compared to the baseline model. On the other hand, a model incorporating only Proposition Type does not improve fit over the baseline model: Model 2 is no better than Model 0. The model incorporating the interaction (Model 4) turns out to have better fit than a model without (Model 3). This suggests that in order to account for the data, we need to account not only the degree of uncertainty, but also for whether or not there was a simple or a compound proposition and, in the latter case, whether the compound proposition presented uncertainty in one or both of its components. An Analysis of

[4]The analysis was conducted in R using the `lme4` package, version 1.1.6 [49]. Model comparisons and estimates of p-values were conducted using the `anova` and `summary` functions in the `lmerTest` package version 2.0.6 [50].

[5]Linear Mixed Effects models combine fixed effects, corresponding to the factors that are explicitly manipulated in an experiment, with random effects, corresponding to factors that may influence the data but are not directly under experimental control. In the present case, the main sources of random effects are subjects – that is, possible differences in responses between participants that go beyond the fixed factors of Uncertainty and Proposition Type – and items – that is, the actual scenarios used, which may themselves have properties that impact responses over and above what is directly manipulated. The random effects part of the model incorporates both intercepts and slopes for both subjects and items, per condition. The random intercepts consider differences between subjects or items when the fixed effects of Uncertainty or Proposition Type are held at their "zero" or default levels. The random slopes incorporate differences between subjects or items as levels of Uncertainty or Proposition Type change.

[6]The Bayesian Information Criterion or BIC is computed as a function of the log-likelihood of a model parameters, given the data, with additional penalties to account for the number of data points required to build the model. A higher BIC suggests a *worse* fit of a model to the data, although BIC is best interpreted as a comparative measure to assess two different models. Note that we also use the -2 log-likelihood measure to compare models directly; this corresponds to the *model* χ^2 value in Table 4, since this quantity is asymptotically χ^2-distributed.

	Fixed effects	BIC	**Model χ^2 ($-2LL$)**
0	Intercept only (baseline)	896.33	–
1	Uncertainty	885.9	17.94^{**} (relative to model 0)
2	Proposition Type	903.31	$0.52, ns$ (relative to model 0)
3	Uncertainty + Proposition Type	884.9	21.56^{**} (relative to model 1) $4.11, ns$ (relative to model 2)
4	Uncertainty × Proposition Type	894.54	27.9^{**} (relative to model 3)
5	Uncertainty × Proposition Type × Language	920.36	$4.23, ns$ (relative to model 4)

Table 4: Model goodness of fit statistics. Models 1, 2 and 3 are compared to the baseline (model 0) to establish the contribution of Properties and Distractors separately. Model 4 is compared to model 3 to establish the contribution of the interaction. Model 5 is compared to Model 4 to establish the role of Language. (*) indicates significantly better goodness of fit at $p < .05$; (**) at $p < .001$.

Variance over this model showed that Proposition Type had only a marginal effect on Ψ ratings ($F(1, 12) = 4.43, p = 0.06$), while Uncertainty exerted a significant main effect ($F(1, 12) = 58.6, p < .001$). The interaction was also significant ($F(1, 11) = 5.23, p < .05$).

When the analysis is extended to incorporate the difference between languages (Model 5), there is no improvement in goodness of fit over Model 4. This is a somewhat surprising result, suggesting that subjective ratings were unaffected by which language participants did the experiment in. This suggests that uncertainty, as measured by possibility and necessity (incorporated in the Ψ measure) might be language independent. Note, however, that this result does not reflect on the differences between languages in their *mapping* from subjective uncertainty to expressions (which are taken into account below).

While the best model so far establishes that Uncertainty exerted a main effect on subjective ratings, with no main effect of type of proposition, it does not determine which levels of Uncertainty actually differed. Did participants tend to rate subjective certainty differently depending on whether there were no uncertain propositions, only one, or both? To establish this, we created versions of Model 4 in which we replaced Uncertainty with a dummy variable obtained by collapsing two levels of Uncertainty together, and comparing them to a third (for example, collapsing the cases where there were 1 or 2 uncertain propositions, comparing them to when there was no uncertainty). This gave rise to three model comparisons, summarised in Table 5.

Comparison	**Model χ^2**
No uncertain prop. vs. 1 uncertain prop.	0
No uncertain prop. vs. 2 uncertain prop.	8^{**}
1 uncertain prop. vs. 2 uncertain prop.	8^{**}

Table 5: Model comparisons for different levels of the Uncertainty factor. All models compared are versions of Model 4, incorporating the full Uncertainty × Proposition Type interaction, collapsing two levels of uncertainty.

The model comparisons suggest that the main effect of Uncertainty was due to subjective

ratings changing mainly in case there were two uncertain propositions, that is, two events in the narrative with fuzzy time intervals. A model that incorporates an Uncertainty variable distinguishing only between this case and the case where there are no uncertain time intervals, or only one, fits the data better.

4.6.1. The impact of temporal granularity

Our scenarios mostly had events with clock times, but in four of the scenarios (see the Appendix), the events had coarser-grained times, specified as years. To illustrate this, consider the experimental scenario in (8) in which participants are asked to evaluate their subjective uncertainty that one event occurred *after* another, and both events have coarse-grained timestamps, given in years.

(8) A team of archaeologists working in Cyprus have uncovered the remains of an old city, built by the Gildeans in the Middle Ages. They suspect that it might be the Holy City mentioned in the Book of Belathon. They know that the city took 20 years to build. They also know that the Book of Belathon took exactly 10 years to write.
Event 1 : The construction of the city started in the middle of 1130.
Event 2 : The writing of the Book of Belathon started in the middle of 1150.
Estimate the certainty of: The writing of the Book of Belathon started after the city was built.

It is possible that participants treated such scenarios differently. One possibility is that humans reason differently about larger time units (such as years), for example by translating them into smaller intervals (e.g. months), or by considering the duration of such events as soft constraints (e.g. in the above example, considering the town as *almost* finished in 1150, or the duration of the construction as being, say, between 18 and 20 years).

	Clock times	**Years**
No uncertainty	0.76 (0.33)	0.82 (0.32)
1 Uncertain proposition	0.53 (0.30)	0.55 (0.28)
2 Uncertain propositions	0.63 (0.30)	0.57 (0.29)

Table 6: Mean Ψ values and standard deviations for different degrees of uncertainty, as a function of temporal granularity

Table 6 displays mean Ψ values for the different uncertainty conditions in the experiment, according to whether times were specified as clock times or years. We tested whether the difference in temporal granularity exerted a statistically reliable impact on subjective uncertainty judgements by incorporating this new variable as an additional fixed effect in Model 4 in Table 4. The new model had an only marginally better fit to the data than Model 4 ($BIC = 915.93$; model $\chi^2 = 4, p = .07$). In other words, we are only able to detect a trend for coarse- vs fine-grained granularities to impact subjective Ψ ratings, a result that we would in any case treat with caution given that granularity was not systematically manipulated as a condition in the experiment, resulting in an uneven distribution of scenarios (4 which involved years, compared to 8 which involved clock times). This is clearly an issue that would warrant more systematic investigation.

English			French			Maltese		
	Π	N		Π	N		Π	N
default (75)	0.98 (0.12)	0.85 (0.27)	default (264)	0.99 (0.1)	0.94 (0.19)	default (114)	1 (0)	0.8 (0.31)
definitely (57)	1 (0)	0.96 (0.12)	sûrement (109)	0.99 (0.09)	0.64 (0.32)	żgur (90)	0.98 (0.11)	0.88 (0.3)
may (152)	0.85 (0.25)	0.07 (0.19)	pourrait (172)	0.78 (0.32)	0.08 (0.21)	jista jkun li (197)	0.91 (0.22)	0.09 (0.21)
must (40)	0.99 (0.06)	0.83 (0.27)	doit (53)	1 (0)	0.71 (0.31)	bilfors li (52)	0.98 (0.14)	0.87 (0.26)
possibly (53)	0.83 (0.29)	0.14 (0.22)	peut-être (183)	0.84 (0.27)	0.05 (0.15)	forsi (62)	0.91 (0.2)	0.14 (0.25)
negation (36)	0.06 (0.18)	0.03 (0.17)	negation (76)	0.12 (0.3)	0.05 (0.2)	negation (45)	0.29 (0.37)	0.04 (0.17)

Table 7: Mean (standard deviation) Π and N values by phrase choice. Frequency of each choice is in parentheses next to the linguistic expression.

4.7. Modelling linguistic expression choice

The second question we outlined at the beginning of this paper concerns the choice of modal expression to express a given degree of uncertainty, In this section, we address this issue in a data-driven fashion. We use the Ψ values to estimate possibility and necessity for each experimental scenario, via equations (6) and (7). Based on these, as well as the type of temporal relation (none, after, before, during), we build a classifier for each language separately, which returns the most likely linguistic expression. We use multinomial logistic regression (equivalent to a maximum entropy approach) for the classification task.

Table 7 displays the mean possibility (Π) and necessity (N) values for each of the three languages we are considering, together with the linguistic expressions chosen by participants (who made their choices following their subjective uncertainty ratings). The table also shows the frequency with which the expressions were selected (see Table 2 for a summary of the possible expressions and a rough equivalence between them).

To obtain an initial view of the accuracy of the classifier, we compared its predicted choice of expression to the actual choice made by experimental participants for each separate observation in the data, based on participants' Π and N values (derived form the Ψ measure). Table 8 shows the proportion of times that a given modal expression was accurately predicted by the model, when compared to the choice made by a human, compared to other possible choices.

A number of interesting observations can be made in relation to the table. In English, where participants select a default (that is, simply asserted, or unmodalised) proposition, the model chooses default roughly the same proportion of the time as a proposition qualifying a relation with *definitely*, suggesting that, as far as possibility and necessity values go, there wasn't much difference between these two cases. A similar situation obtains in French and Maltese in the second row of their respective sub-tables, where for the human choice of *sûrement* or *żgur*, the model is more likely to select the default case, followed by the correct choice. Interestingly, the expression *possibly* in English and its Maltese equivalent are never selected by the model: in the majority of cases where humans select this expression, the model's own choice is *may*, suggesting once again that there is a close relationship in subjective Π and N values between these two. By contrast, the classifier trained on French data selects the equivalent of *possibly* more frequently.

4.8. Summary

The experiment reported in this section addressed two empirical questions. First, using the Ψ scale to elicit possibility and necessity ratings, we showed that subjective uncertainty in temporal

	default	definitely	may	must	possibly	negation
default	0.43	0.43	0.11	0.03	0.00	0.01
definitely	0.14	0.82	0.04	0.00	0.00	0.00
may	0.07	0.00	0.91	0.01	0.00	0.01
must	0.43	0.35	0.13	0.10	0.00	0.00
possibly	0.04	0.00	0.89	0.02	0.00	0.06
negation	0.00	0.03	0.06	0.00	0.00	0.92

(a) Model predictions for English.

	default	sûrement	pourrait	doit	peut-être	negation
default	0.94	0.02	0.03	0.00	0.01	0.00
sûrement	0.50	0.31	0.05	0.00	0.14	0.00
pourrait	0.03	0.07	0.27	0.00	0.56	0.07
doit	0.64	0.15	0.11	0.00	0.09	0.00
peut-être	0.01	0.03	0.27	0.00	0.64	0.04
negation	0.04	0.03	0.03	0.00	0.04	0.87

(b) Model predictions for French.

	default	żgur	jista jkun li	bilfors li	forsi	negation
default	0.65	0.13	0.13	0.09	0.00	0.00
żgur	0.53	0.22	0.09	0.14	0.00	0.01
jista jkun li	0.06	0.01	0.88	0.01	0.00	0.05
bilfors li	0.46	0.25	0.08	0.19	0.00	0.02
forsi	0.10	0.02	0.84	0.00	0.00	0.05
negation	0.04	0.00	0.31	0.00	0.00	0.64

(c) Model predictions for Maltese.

Table 8: Predicted choice of epistemic modal phrase, against human choices. Diagonals in each subtable are proportions where model and human predictions coincide.

relations is impacted by the degree of uncertainty associated with the time intervals of individual events, and to a lesser extent by the type of temporal relation under consideration. Crucially, this turned out to be largely independent of which language a narrative is presented in. We also described separate language-based classifiers, based on multinomial logistic regression, which predicted the choice of epistemic modal expression given the Π and N values for a specific scenario, and the type of temporal relation involved. Fitting the model to the actual observations in the data sheds some light on cross-linguistic differences in the type of expression people are likely to use to express their degree of certainty about a temporal relation.

5. Evaluation of the model

We now turn to an evaluation of the formalism for temporal reasoning based on fuzzy temporal intervals presented in Section 3.

5.1. Data

For the purposes of the evaluation, a different dataset was generated using the same methodology as that described in Section 4. The same scenarios were presented to new subjects, from whom Ψ ratings and expression choices were elicited. The procedure was identical to the previous experiment and materials were counterbalanced in the same way. There were 13 participants for each of the three languages.

5.2. Evaluating the temporal reasoning formalism

We begin by evaluating the temporal formalism, by comparing the human subjective uncertainty judgements to those predicted by the model. To achieve this, we computed the Π and N values for each scenario, as follows. First, we translated each scenario into its component event representation, along the lines shown in Figure 3, and then used the reasoning engine described in Section 3 to compute the relevant values. This computation relied on three assumptions:

1. if a scenario stated that an event occurred at a specific time (or within an interval), the event was represented with that time or interval as its start time;
2. over a given uncertain interval, the possibility distribution for an event was uniform, that is, if an event was stipulated as having started between t_0 and t_1, it was equally possible/necessary during any subinterval of $[t_0, t_1]$.
3. when an event duration was only expressed in terms of maximum duration (e.g., "not more than 10 minutes") the possibility distribution was translated by the intuitive fuzzy set $\{MAX/2, MAX-1, MAX/2, 2, 1\}$ where MAX is the maximum duration. This corresponds to the fact that the event at least had a duration (that is, a duration of 0 is not possible) and that the duration is more likely to be around the last part of the interval (e.g., 5 to 10 minutes is completely possible, but less than five minutes is less possible). Finally the MAX value cannot exceed .5, since it has been observed that, for humans, it is unclear whether the MAX is included in an interval or not.

From the Π and N values computed by FuzzyTIME, the value of Ψ can be reconstructed using $\Psi(P) = \frac{1}{2}[\Pi(P) + N(P)]$. These machine-generated values were subsequently correlated to the mean Ψ value obtained from participant judgements in each experimental condition. Table 9 summarises the correlations for each language, and overall.[7]

	Pearson's r	p-value
French	.46	$< .001$
English	.57	$< .001$
Maltese	.5	$< .001$
Overall	.67	$< .001$

Table 9: Correlations between computed and elicited Ψ judgements. P-values indicate the likelihood that the degree of covariation denoted by r is due to chance.

As Table 9 suggests, all correlations were positive and highly significant, and higher when averaged over all languages. The value of $r = .67$ for the overall correlation suggests that the FuzzyTIME predictions can account for approximately ($.67^2 =$) roughly 45% of the variance in the human Ψ ratings. While this is not perfect, it does suggest that the model is on the right track.

[7] We use Pearson's r as a measure of correlation, which is estimated as a function of the co-variance of two numerical variables. The closer r is to 1 or -1, the stronger the correlation.

5.3. Evaluating the phrase choice classifier

We next evaluate the classifiers described in the previous section for each language, by running them over the evaluation data and comparing their predicted phrase to the one selected by humans in the evaluation study. Table 10 displays proportions of accurate classifications for each language. We compare this to a baseline model which simply selected the majority classification in each case, and a second baseline model which made a random selection of phrase.

	English	**French**	**Maltese**
Model Accuracy	56%	55%	52%
Majority class baseline	30%	30%	32%
Random baseline	16%	25%	26%

Table 10: Accuracy statistics for the phrase-choice model, applied to the evaluation data

As the table shows, the models predicting language-specific modal expressions from Π and N outperform both baselines by a considerable margin. Nevertheless, model accuracy does not exceed 56%. While this seems an acceptable performance, it is clear that other factors play a role in epistemic modal expression selection. As shown at the end of Section 4, human choices seem to be motivated by different factors depending on the language they are using. Furthermore, certain choices made by the model seem to reflect a close proximity between alternative expressions – this was shown to be the case between, say, unmodalised ('default') cases and cases corresponding to the expression *definitely*, as well as the cases corresponding to *may* and *possibly*. Clearly, much more research is required into the differences between such expressions and into how an NLG system can make such choices in a manner that approximates human performance.

6. Conclusions

This paper addressed the problem of temporal uncertainty and its expression in Natural Language Generation systems that summarise raw data in which event times play a role. We have argued that such systems should convey uncertainty to the reader, exploiting the mechanisms made available in natural language, especially epistemic modal expressions. We addressed these issues formally, through a model based on Fuzzy Temporal Constraint Networks, and empirically, through experiments involving human participants.

One question is whether a specific model of fuzzy temporal representation based on Possibility Theory can make predictions that correlate with human uncertainty judgements. Our results suggest that this is likely to be the case. A second question is to what extent the uncertainty estimates of such a model can be mapped to epistemic modal expressions, in order to express uncertainty in the data and convey it to the reader, if it exists. We have used our experimental data to propose a classification model that maps from subjective uncertainty to expressions with epistemic modal force, in three different, and typologically distinct, languages. The results on a new evaluation dataset are promising.

The work reported here also opens up a number of avenues for future research, both from a methodological, and a theoretical point of view.

Our data was obtained through controlled experiments with human subjects, using linguistic stimuli in the form of narratives. The experimental approach was partly motivated by a lack of data that is adequate for the task at hand, namely, to quantify a subjective notion of uncertainty

within a formally precise framework, and to map this uncertainty to linguistic expressions which could convey it to a reader. Nevertheless, the use of linguistic stimuli may have been taxing on participants, who were required to reason about quite complex scenarios. An extension of this work would benefit from a consideration of non-linguistic stimuli (for example, animations), while eliciting descriptions of stimuli more freely from subjects.

Our analysis also uncovered a number of possible factors that could influence both human subjective uncertainty and phrase choice. These include the temporal granularity at which events are specified (e.g. years vs. clock times), as well as apparent cross-linguistic differences in the way propositions are modalised. A more precise picture of the relationship between the formal possibility-theoretic framework used here, and linguistic descriptions of temporally uncertain data, would call for in-depth investigation into these issues.

Finally, we have restricted attention to scenarios involving temporal relations between two events. More complex relationships between three or more events, would raise new challenges computationally, insofar as possibility/necessity values for a given case have to take into account more than just a single pairwise relationship. They would also increase the complexity of the linguistic description task, since the options to describe such scenarios would likely increase, and extend well beyond the distinction between simple propositions, and propositions with a temporal connective.

In summary, this paper sought to bridge between two areas, that of possibility theory, and that of language generation. The empirical perspective taken in this paper was exploited in two directions: first, with a view to analysing human linguistic choices, as well as the conditions which influence their subjective uncertainty about time and temporal relations; and second, with a view to evaluating the temporal reasoning formalism based on Possibility Theory, and the classifier trained on human data to select from alternative modal expressions. The outcomes of this study are encouraging, and suggest that the two fields would benefit from increased cross-fertilisation.

Appendix: List of scenarios in the experiments

The following are the thirteen scenarios constructed for the experiments. In each case, there is a short background narrative followed by two key events, either or both of which can have a precise or fuzzy temporal interval (shown in parentheses below). Each of these scenarios was used in either of the three languages of the experiments, having been translated from the original English by a native speaker.

1. A bank robbery occurred yesterday afternoon. An investigator is trying to reconstruct the scene from eye-witness reports. He knows for certain that the robbers were inside the bank for no more than 45 minutes. He also knows for certain that the police took exactly 30 minutes to arrive on the scene after being alerted. He has also interviewed some eye-witnesses. Here is what they said:
The robbers entered the bank at 16:00 (some time between 16:00 and 16:30).
The police were alerted at 16:15 (some time between 16:15 and 16:45).

2. A team of archaeologists working in Cyprus have uncovered the remains of an old city, built by the Gildeans in the Middle Ages. They suspect that it might be the Holy City mentioned in the Book of Belathon. They know that the city took 20 years to build. They also know that the Book of Belathon was written over a period of 10 years.

The construction of the city started in 1130 (some time between 1130 and 1145).
The writing of the Book of Belathon started in 1150 (some time between 1145 and 1150).

3. Percy Maxwell was famous both as a poet and as a politician. However, relatively little is known about his life. Historians know for certain that his most famous poem, The Agathon, took him 15 years to write. He was also married, but got divorced after 10 years.
Maxwell began writing the Agathon in 1650 (some time between 1650 and 1660).
He got married in 1660 (some time between 1655 and 1665).

4. Stanley was late getting to the pub and missed his rendezvous with Mirella and Jake. The barman, an old friend, told him that Mirella had certainly not stayed longer than 30 minutes. Jake had stayed for 45 minutes and then left.
Mirella had arrived at 6pm (some time between 6:00 and 6:30).
Jake had arrived at 6:15 (some time between 6:15 and 6:45).

5. There was an accident at the airport. A plane's engine caught fire. We know for certain that it takes 5 minutes for a fire to trigger the plane's fire alarm. We also know for certain that the plane was in the air for exactly 25 minutes before it made an emergency landing.
The plan left the ground at 10:05 (some time between 10:05 and 10:25).
The fire started at 10:10 (some time between 10:10 and 10:20).

6. Historians are trying to identify the date of creation of an old manuscript and a painting, both from the Victorian era. They know for certain that the manuscript was written by a monk, who took five years to complete it and then died immediately after. They also know that it took the artist three years to finish the painting.
The monk started work on the manuscript in 1815. (some time between 1815 and 1820).
he artist started work on the painting in 1816. (some time between 1816 and 1819).

7. A young girl has been reported missing. She had been shopping with her mother when she wandered off. She was seen by two people. One of them saw her near the town hall. She is certain that she was there for 5 minutes at most. Another person saw her talking to a woman, but can't remember where, though he is sure that they chatted for 15 minutes at most.
The girl was seen at the town hall at 11:15 (some time between 11:15 and 11:25).
She was seen talking to the woman at 11:18 (some time between 11:18 and 11:35).

8. Alison and Jim bought a very old house. It had been built in the eighteenth century. The east wing had taken 3 years to build, while the west wing took 2 years.
The east wing was started in 1760 (some time between 1760 and 1761).
The west wing was started in 1762 (some time between 1762 and 1763).

9. Judith and Albert were meant to meet in the hotel lobby, but they just missed each other. Judith took the lift from the tenth floor, and it took her three minutes to get to the lobby. Albert took the stairs from the fifth floor, and it took him four minutes to get to the lobby.
Judith left her room at 9pm (some time between 9pm and 9:02pm).
Albert left his room at 9:01 pm (some time between 9:01 pm and 9:03 pm).

10. Police are investigating a shooting incident that occurred in a suburb of London. Some men were observed entering a building. They were in there for exactly 15 minutes. Several shots were fired. The shooting continued for 10 minutes.
The men entered the building at 11 pm (some time between 11 pm and 11:05).
The shooting started at 11:02 pm (some time between 11:02 and 11:07).

11. Sally's husband was worried about her. She had gone hillwalking that morning and a thunderstorm had occurred. Sally was on the hill for two hours. The thunderstorm lasted for one hour.
Sally started her walk at 10:15 (some time between 10:15 and 11:15).
The thunderstorm started at 11:05 (some time between 11:05 and 11:20).

12. There was a riot at the football stadium. It took the police an hour to disperse the rioters. Meanwhile, play continued on the field. The second half lasted for 50 minutes because the referee granted the players 5 minutes of injury time.
The riot started at 4pm (some time between 4pm and 4:10pm).
The second half started at 4:05 (some time between 4:05 and 4:15).

13. Aline needed to catch a train to Edinburgh. This train journey lasts exactly two and a half hours. However, she had to attend a meeting first, which lasted an hour.
Aline's meeting started at 10:00 (some time between 10:00 and 10:05).
The train to Edinburgh left at 11:05 (some time between 11:05 and 11:15).

References

[1] A. Gatt, F. Portet, E. Reiter, J. Hunter, S. Mahamood, W. Moncur, S. Sripada, From data to text in the neonatal intensive care unit: Using NLG technology for decision support and information management, AI Communications 22 (2009) 153–186.

[2] P. Terenziani, R. T. Snodgrass, A. Bottrighi, M. Torchio, G. Molino, Extending temporal databases to deal with telic/atelic medical data., in: Proceedings of the 10th Conference on Artificial Intelligence in Medicine (AIME 05), 2005.

[3] E. Hoffer, M. Feldman, R. Kim, K. Famiglietti, G. Barnett, Dxplain: Patterns of use of a mature expert system., in: AMIA Annu Symp, 2005, pp. 321–325.

[4] E. Reiter, R. Dale, Building Natural Language Generation systems., Cambridge University Press, 2000.

[5] E. Reiter, Natural Language Generation, in: A. Clark, C. Fox, S. Lappin (Eds.), Handbook of Computational Linguistics and Natural Language Processing, no. May 2009, Wiley, Oxford, 2010, pp. 574–598.

[6] E. Reiter, An architecture for data-to-text systems, in: Proceedings of the 11th European Workshop on Natural Language Generation (ENLG'07), Association for Computational Linguistics, Stroudsberg, PA, 2007, pp. 97–104.
URL http://dl.acm.org/citation.cfm?id=1610163.1610180

[7] E. Goldberg, N. Driedger, R. I. Kittredge, Using Natural Language Processing to Produce Weather Forecasts, IEEE Expert (1994) 45–53.

[8] E. Reiter, S. Sripada, J. Hunter, J. Yu, I. Davy, Choosing words in computer-generated weather forecasts, Artificial Intelligence 167 (1-2) (2005) 137–169. doi:10.1016/j.artint.2005.06.006.
URL http://linkinghub.elsevier.com/retrieve/pii/S0004370205000998

[9] T. Bethem, J. Burton, T. Caldwell, M. Evans, R. Kittredge, B. Lavoie, J. Werner, Generation of Real-time Narrative Summaries for real-time water levels and meteorological observations in ports, in: Proceedings of the Fourth Conference on Artificial Intelligence Applications to Environmental Sciences (AMS'05), 2005, pp. 1–4.

[10] M. Kahn, L. Fagan, L. Sheiner, Combining physiologic models and symbolic methods to interpret time-varying patient data, Methods of information in medicine 30 (1991) 167–178.

[11] D. Hüske-Kraus, Text generation in clinical medicine–a review., Methods of information in medicine 42 (1) (2003) 51–60. doi:10.1267/METH03010051.
URL http://www.ncbi.nlm.nih.gov/pubmed/12695796

[12] M. D. Harris, Building a large-scale commercial NLG system for an EMR, in: Proceedings of the Fifth International Natural Language Generation Conference (INLG '08), Association for Computational Linguistics, Morristown, NJ, USA, 2008, pp. 157–160. doi:10.3115/1708322.1708351.
URL http://portal.acm.org/citation.cfm?doid=1708322.1708351
[13] J. Yu, E. Reiter, J. Hunter, C. Mellish, Choosing the content of textual summaries of large time-series data sets, Natural Language Engineering 13 (01) (2006) 25. doi:10.1017/S1351324905004031.
URL http://www.journals.cambridge.org/abstract_S1351324905004031
[14] M. Molina, A. Stent, E. Parodi, Generating Automated News to Explain the Meaning of Sensor Data, in: J. Gama, E. Bradley, J. Hollmén (Eds.), Proceedings of IDA 2011, Springer, Berlin and Heidelberg, 2011, pp. 282–293.
[15] F. Portet, E. Reiter, A. Gatt, J. Hunter, S. Sripada, Y. Freer, C. Sykes, Automatic generation of textual summaries from neonatal intensive care data, Artificial Intelligence 173 (7–8) (2009) 789–816.
[16] J. Hunter, Y. Freer, A. Gatt, E. Reiter, S. Sripada, C. Sykes, Automatic generation of natural language nursing shift summaries in neonatal intensive care: Bt-nurse, Artificial Intelligence in Medicine 56 (3) (2012) 157–172.
[17] E. Reiter, A. Gatt, F. Portet, M. Van Der Meulen, The importance of narrative and other lessons from an evaluation of an NLG system that summarises clinical data, in: Proceedings of the 5th International Conference on Natural Language Generation (INLG-08), Salt Fork, Ohio, United States, 2008, pp. 147–153.
[18] A. Kratzer, The notional category of modality, in: H.-J. Eikmeyer, H. Rieser (Eds.), Words, Worlds and Contexts: New Approaches to Word Semantics, Walter de Gruyter and Co., Berlin, 1981.
[19] A. Gatt, F. Portet, If it may have happened before, it happened, but not necessarily before, in: Proceedings of the 13th European Workshop on Natural Language Generation, (ENLG'11), 2011, pp. 91–101.
[20] F. Palmer, Mood and modality, 2nd Edition, Cambridge University Press, Cambridge, 2001.
[21] M. Vanhove, C. Miller, D. Caubet, The grammaticalisation of modal auxiliaries in maltese and arabic vernaculars of the mediterranean area, in: B. Hansen, F. de Haan (Eds.), Modals in the languages of Europe, Mouton, Berlin, 2009.
[22] N. Evans, S. Levinson, The myth of language universals: Language diversity and its importance for cognitive science, Behavioral and Brain Sciences 32 (2009) 429–492.
[23] R. Klabunde, Lexical choice for modal expressions, in: Proceedings of the 11th European Workshop on Natural Language Generation (ENLG-07), 2007.
[24] R. Klabunde, When must should be chosen, in: Proceedings of the 10th European Workshop on Natural Language Generation (ENLG-05), 2005.
[25] P. Portner, Modality, Oxford University Press, Oxford, 2009.
[26] A. Papafragou, Inference and word meaning: The case of modal auxiliaries, Lingua 105 (1998) 1–47.
[27] E. E. Sweetser, From etymology to pragmatics, Cambridge University Press, Cambridge, 1990.
[28] J. F. Allen, Maintaining knowledge about temporal intervals, Communications of the ACM 26 (11) (1983) 832–843.
[29] R. Dechter, I. Meiri, J. Pearl, Temporal constraint networks, Artificial Intelligence 49 (1991) 61–95.
[30] M. Vilain, H. Kautz, P. van Beek, Constraint propagation algorithms for temporal reasoning: A revised report, in: D. S. Weld, J. de Kleer (Eds.), Readings in Qualitative Reasoning about Physical Systems, Morgan, 1987, pp. 373–381.
[31] J. Palma, J. M. Juarez, M. Campos, R. Marín, Fuzzy theory approach for temporal model-based diagnosis: An application to medical domains, Artificial Intelligence in Medicine 38 (2) (2006) 197–218.
[32] F. Gao, S. Sripada, J. Hunter, F. Portet, Using temporal constraints to integrate signal analysis and domain knowledge in medical event detection, in: 12th Conference on Artificial Intelligence in Medicine (AIME 09), Vol. 5651 of LNAI, 2009, pp. 46–55.
[33] L. Zhou, G. Hripcsaka, Temporal reasoning with medical data — a review with emphasis on medical natural language processing, Journal of Biomedical Informatics 40 (2) (2007) 183–202.
[34] A. Artikis, A. Skarlatidis, F. Portet, P. G., Logic-based event recognition, Knowledge Engineering Review 27 (4) (2012) 469–506.
[35] L. A. Zadeh, Fuzzy sets, Information and Control 8 (3) (1965) 338–353.
[36] S. Barro, R. Marín, J. Mira, A. Paton, A model and a language for the fuzzy representation and handling of time, Fuzzy Sets and Systems 61 (1994) 153–175.
[37] L. Vila, L. Godo, On fuzzy temporal constraint networks, Mathware and soft computing 3 (1994) 315–334.
[38] M. Campos, A. Cárceles, J. Palma, R. Marín, A general purpose fuzzy temporal information management engine, in: EurAsia-ICT 2002, 2002, pp. 93–97.
[39] R. Marín, M. A. Cárdenas, M. Balsa, J. L. Sanchez, Obtaining solutions in fuzzy constraint networks, International Journal of Approximate Reasoning 16 (3-4) (1997) 261 – 288.
[40] D. Dubois, H. Allel, H. Prade, Fuzziness and uncertainty in temporal reasoning, Journal of Universal Computer Science 9 (9) (2003) 1168–1194.
[41] F. Portet, A. Gatt, Towards a possibility-theoretic approach to uncertainty in medical data interpretation for text generation, in: D. Riano, A. ten Teije, S. Miksch, M. Peleg (Eds.), Knowledge Representation for Health-Care:

Data, Processes and Guidelines, LNAI 5943, Springer, Berlin and Heidelberg, 2010.
[42] E. Raufaste, R. D. S. Neves, C. Mariné, Testing the descriptive validity of possibility theory in human judgments of uncertainty, Artificial Intelligence 148 (1-2) (2003) 197–218.
[43] N. Cowan, The magical number 4 in short-term memory: A reconsideration of mental storage capacity, Behavioral and Brain Sciences 24 (2001) 87–114.
[44] A. Baddeley, Working Memory, Thought, and Action, Oxford University Press, Oxford, 2007.
[45] P. C. Wason, Shapiro, Natural and contrived experience in a reasoning problem, Quarterly Journal of Experimental Psychology 23 (1971) 63–71.
[46] J. Evans, Bias in human reasoning: causes and consequences, Psychology Press, 1990.
[47] C. Emmott, Narrative Understanding, Oxford University Press, Oxford, 1999.
[48] A. Sanford, C. Emmott, Mind, Brain and Narrative, Cambridge University Press, Cambridge, 2012.
[49] D. Bates, M. Maechler, B. Bolke, lme4: Linear mixed-effects models using S4 classes. (2014).
URL http://cran.r-project.org/web/packages/lme4/index.html
[50] A. Kuznetsova, P. B. Brockhoff, R. H. B. Christensen, lmerTest: Tests for random and fixed effects for linear mixed effect models (2014).
URL http://cran.r-project.org/web/packages/lmerTest/index.html
[51] D. J. Barr, R. Levy, C. Scheepers, H. J. Tily, Random effects structure for confirmatory hypothesis testing: Keep it maximal., Journal of memory and language 68 (3) (2013) 255–278.