If it may have happened before, it happened, but not necessarily before

Albert Gatt
Institute of Linguistics
University of Malta
Tal-Qroqq, Msida MSD2080
Malta
`albert.gatt@um.edu.mt`

François Portet
LIG UMR 5217, UJF-Grenoble 1
Grenoble INP / UPMF-Grenoble 2 / CNRS,
Laboratoire d'Informatique de Grenoble
Grenoble, F-38041, FRANCE
`francois.portet@imag.fr`

Abstract

Temporal uncertainty in raw data can impede the inference of temporal and causal relationships between events and compromise the output of data-to-text NLG systems. In this paper, we introduce a framework to reason with and represent temporal uncertainty from the raw data to the generated text, in order to provide a faithful picture to the user of a particular situation. The model is grounded in experimental data from multiple languages, shedding light on the generality of the approach.

1 Introduction

Natural Language Generation (NLG) systems which take raw data as input often need to transform it by performing operations such as inference, abstraction or approximation. However, in many domains, input data is riddled with uncertainty or inaccuracy. For example, a patient database may contain records of interventions which were entered well after they actually occurred (Gatt et al., 2009). This problem is particularly acute in systems where the temporal dimension of the data is important; it is exacerbated by the lack of a principled way of handling temporal information in existing database management systems (Terenziani et al., 2005).

Temporal uncertainty – that is, uncertainty about the precise time at which an event occurred – can affect NLG systems at the data processing and document planning stages, since it affects temporal and/or causal relationships between events. It also impacts microplanning and realisation, since decisions must be made as to whether a proposition is

to be simply asserted or modalised to express some degree of epistemic (un)certainty. Simply asserting a proposition will normally give rise to the presupposition that the state of affairs described is known for certain (Karttunen, 1972); conversely, modalising the proposition impacts its truth conditions (Papafragou, 2006).

In this paper, we argue that temporal uncertainty should be explicitly communicated, and we focus on the use of modalised propositions to acheive this[1], taking a multilingual perspective. Our aim is to address two empirical questions. The first concerns the (non-linguistic) 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 'subjective' 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 a formalism to handle this, showing that its predictions have a good correspondence to human intuitions. Our second question concerns the way in which modal expressions can be *grounded* in subjective uncertainty arising from raw data. We describe an experimental design 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

---

[1]In what follows, our use of the term 'modality' refers to the semantic or 'notional' category (Kratzer, 1981). As Kratzer argues, this can be expressed in a variety of ways, ranging from modal auxiliaries to adverbs of possibility, among others.

close (insofar as they are European), are typologically diverse. In this way, we 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), and the experiment and results (Section 4). We conclude in Section 5 with some pointers to future work.

2 Epistemic uncertainty in language

The expression of uncertainty is usually achieved through modal expressions, which are concerned with the degree of possibility or necessity associated with a particular proposition. Modality, which is often associated (and in some languages, conflated) with the category of Irrealis, can be characterised in terms of assertion (Palmer, 2001): an unmodalised proposition is simply asserted (thereby presupposing certainty 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.

We are primarily interested in how the resources that a language makes available to express epistemic modality can be harnessed to express temporal uncertainty in data-to-text systems, thus avoiding misleading the reader. While the importance of this problem has been pointed out in recent work (Portet et al., 2009; Gatt et al., 2009), modality lacks a principled treatment in NLG (but see Klabunde (2007)). As Klabunde notes, NLG systems which use modals in their output (Elhadad, 1995; Reiter et al., 2003) do not seem to select these expressions in a principled way. The following example illustrates some of the difficulties in dealing with epistemic modality, especially from a cross-linguistic perspective:

(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 (Palmer, 2001). 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*
            the.fsg police   may.3pl/must.3pl   have   be.3sg.ps
            sur les   lieux   avant que les   voleurs
            on   def.pl scene.pl before   the.pl robber.pl
            quittent   la   banque.
            leave.3pl.ps the.fsg bank
            'The police **may/must** have been on the scene before the robbers left the bank.'

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

The above example suggests certain similarities between English and French, despite their different genetic classification (Anglo-Saxon vs. Romance). The difficulties increase when other language families are considered. We will also consider a European language which comes from a third language family, namely Maltese (Semitic), where the modal auxiliaries that have been identified (Vanhove et al., 2009) 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 this is only possible with the imperfective form and is more frequently rendered in a construction involving clausal subordination using *li* ('that'), a form that is also commonly used with adverbs like *bilfors* (e.g. *bilfors li*; lit. 'by force that', i.e. 'definitely') and *żgur* ('certainly') (3a). One adverbial that normally occurs without explicit subordination of the matrix VP is the Romance-derived *forsi* ('maybe/perhaps') (3b). However, current descriptive work on these modals does not give a clear picture of the difference in the distribution of these expressions and suggests that some of them may be highly restricted in their use.

(3)   (a) *Il-pulizija* **jista' jkun/bilfors/żgur**   *li*   *kienu*
       the-police   could be/definitely/certainly that be.pl.ps
       fuq   *ix-xena*   *qabel ma* *l-ħallelin*   *telqu*
       on   the-scene   before   the-robber.pl leave.pl.ps
       mill-bank.
       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*
       the-police possibly be.3pl.ps on   the-scene
       qabel ma *l-ħallelin*   *telqu*   *mill-bank.*
       before   the-robber.pl 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 (Evans and Levinson, 2009). 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 Klabunde (2007), who 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 (Klabunde, 2005; Klabunde, 2007). Klabunde's approach is based on the influential possible worlds framework proposed by Kratzer (Kratzer, 1977; Kratzer, 1981; Portner, 2009), 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.

Most semantic work on modality has been based on this framework (but see Papafragou (1998) for a relevance-theoretic account, and Sweetser (1990) for a cognitive-functionalist account). Neither of these theories is 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. Nor is it immediately obvious, were the problem amenable to such a treatment, that this is the most cognitively plausible or computationally tractable way of representing uncertainty, relying as it does on an exhaustive consideration of alternative situations (Johnson-Laird, 1978). In the following section, we consider an alternative proposal.

3 Temporal representation and reasoning

The formalism used to represent and reason with events and relations between events is based on the Metric Temporal Constraint Network (TCN) (Dechter et al., 1991) approach.

This approach differs from purely qualitative ap-

proaches — such as the one based on Allen's thirteen mutually exclusive binary relations (Allen, 1983)— as it considers only metric-based temporal relations (e.g., 'Mary left 10 minutes before James arrived' as opposed to 'Mary left before James arrived') and represents events as time-points rather than intervals 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 recuding the expressiveness of the interval relations (see Vilain et al. (1987)). Moreover, there are numerous algorithms to compute the consistency of a TCN network efficiently, depending on the allowed experessivity, 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 (Palma et al., 2006; Gao et al., 2009). For more information about temporal reasoning and the aforementioned formalisms the reader is refereed to (Zhou and Hripcsaka, 2007; Artikis et al., 2010).

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. An instantaneous event or **event** a is a tuple $\langle t, o \rangle$, where $t \in \mathbf{N}$ and $o \in \mathcal{O}$. t is the known date of occurrence of the event and o represents some structured data corresponding to this event (e.g. database record, inference, user input). Among other things, o can correspond to a type (concept) in a knowledge repository such as an ontology $\mathcal{O}$. A durative event or **interval** A is a tuple $\langle as, ae, c, o \rangle$, where as (resp. ae) is an instantaneous event representing the start (resp. end) of the durative event, c is a numerical constraint such that $ae - as \in (0, c]$ and o is the description of the durative event.

Briefly, a TCN $\mathcal{N}$ consists of a set of instantaneous events (a, b, c) with constraints between them. Each constraint T between a and b is repre-

sented by a set of binary constraints ($\{I_1, \dots, I_n\} = \{[t_{s1}, t_{e1}], \dots, [t_{sn}, t_{en}]\}$) that represent the temporal knowledge about a situation. For instance, the set of facts in example (1), can be represented by the TCN depicted in Figure 1 where all durative events are translated into pairs of events (e.g. 'were inside the bank' $\rightarrow$ 'robbers enter' and 'robbers leave') and all temporal relations are translated into binary temporal constraints (e.g., 'for no more than 45 minutes' $\rightarrow [1, 45]$). This also applies to absolute times, which are represented with respect to the origin of the day.
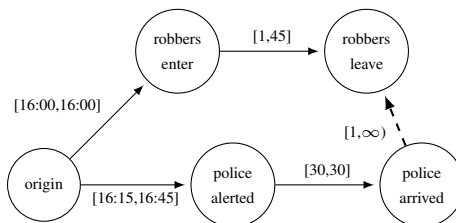

Figure 1: Robbers example represented as a TCN.

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., $t_{s1} \leq b - a \leq t_{e1} \vee \cdots \vee t_{sn} \leq b - a \leq t_{en}$). Briefly, this consists in applying algorithms that solve the shortest path problem to generate the minimal network (i.e., the network with the tightest constraints). If one constraint is not satisfied then no solution exists and the network is inconsistent. For instance, if one wants to test the assertion *The police were on the scene before the robbers left the bank*, this constraint can be integrated into the network (before $\rightarrow [1, \infty)$; see the dashed edge in Figure 1) and the consistency checking algorithm will find no solution, because 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. 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. To better capture these intuitions, it is possible to represent each temporal constraint as a fuzzy set (Zadeh, 1965).

There are several implementations of Fuzzy Temporal Constraint Networks (FTCNs) (Barro et al., 1994; Vila and Godo, 1994; Campos et al., 2002). We will focus on the one implemented in the Fuzzy-TIME engine (Barro et al., 1994; Campos et al., 2002). 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. In this approach, a binary constraint between two events is defined by a normalised, unimodal possibility distribution π which restricts the temporal distance between two events. Recall that in possibility theory (Dubois et al., 2003), 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.

An example FTCN is represented in Figure 2 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.* 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).

In FTCN, the solutions to the network can satisfy the constraints only to a certain degree σ, given
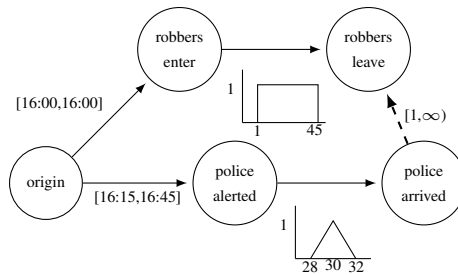

Figure 2: Robbers example represented as a FTCN.

that temporal constraints may be fuzzy. In Fuzzy-TIME, 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) (Barro et al., 1994). For instance, incorporating the assertion *The police were on the scene before the robbers left the bank.* with $\Delta_t = 1$ minute leads to a network consistent with only .5 possibility and 0 necessity (because the 'after' relation is completely possible).

This model therefore offers us the possibility 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.

4 Experiment

In this section, we describe an experiment whose aims were (1) to validate the possibility-theoretic formalism against human data, by comparing uncertainty computations to human subjective evaluations based on the same scenarios; (2) to map subjective certainty judgements to the classes of modal expressions in French, Maltese and English introduced in Section 2, thereby also testing whether the formalism itself can adequately capture subjective uncertainty judgements by speakers of different languages. The experiment replicated the one reported by Portet and Gatt (2010), with some differences in the choice of materials, and with the crucial differ-

ence that it was carried out on three groups of native speakers of the three languages under consideration. Furthermore, we go beyond their analysis in comparing the possibility-theoretic formalism to human judgements.

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

Table 1: Modal expressions used in the experiment.

Design and procedure   The experiment exposed participants to scenarios such as those in example (1) through a web interface; this is partially displayed in Figure 3. 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 fuzzy 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:

1. Judgement: Participants were given a 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 3, top). To elicit these subjective certainty judgements, we used a slider representing the Ψ-scale developed by Raufaste et al. (2003). 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.

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

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

2. Expression choice: For each scenario, participants were also presented with a list of 6 different versions of the proposition they had judged in random order and asked to choose the one that they felt best reflected their degree of certainty (Figure 3, 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 1. 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 seek to address.

The experimental scenarios represented combinations of two within-participants factors. *Uncertainty* (3 levels) manipulated the amount of temporal uncertainty in scenario, where either both key events were given an exact time (e.g. *at 16:00*), or one had a fuzzy temporal interval (e.g. *between 16:00 and 16:45*) or both did. *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 describing a temporal relation between them using one of the 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. There was also a third, between-groups factor, namely Language (Maltese/English/French). Thus, our experiment had a mixed 3 (Uncertainty) $\times 4$ (Proposition Type) $\times 3$ (Language) design.

Materials and participants   Thirteen scenarios were constructed; each one had a version in English, Maltese and French. Within each language, each one had 13 different versions corresponding to
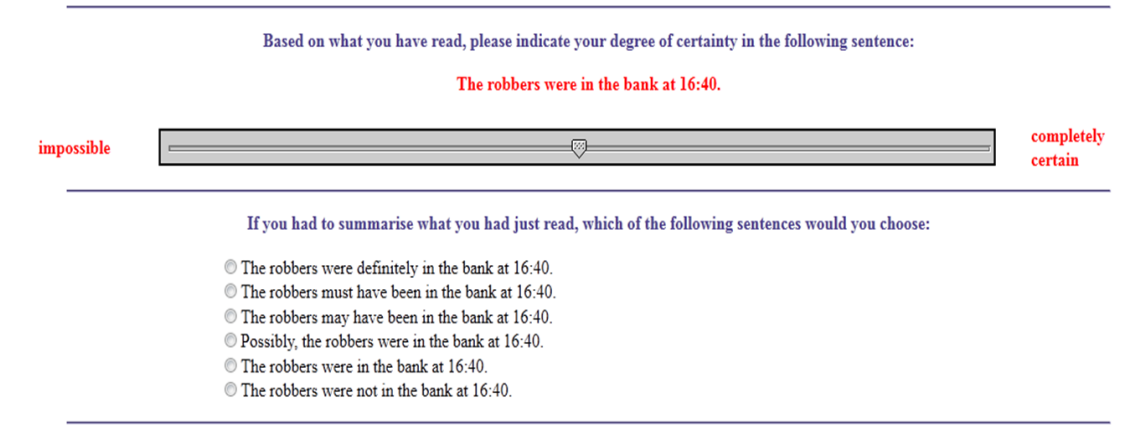
Figure 3: Partial screenshot of the experiment interface

each of the 13 conditions. 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. The present analysis is based on data from 3 different groups of 13 native speakers of each language. Within each group, each participant did one of the versions of the experiment.

4.1 Results

We first test the effects of Uncertainty and Proposition Type on subjective uncertainty judgements using the Ψ scale and compare the subjective judgements made by experimental participants to the output of the reasoning engine on the same scenarios. We then attempt to model statistically the mapping from subjective uncertainty to choice of linguistic expressions.

4.1.1 Subjective uncertainty

Table 2 summarises the mean Ψ ratings overall and within each language, as a function of the different levels of Proposition Type. At a glance, there is a clear tendency for subjective certainty to decrease as scenarios introduce more temporal uncertainty, as expected. However Proposition Type seems to affect ratings less drastically. To test these intuitions, we used a linear mixed effects analysis, with our three factors (Uncertainty, Proposition Type and Language) as fixed effects, and participants and items as random effects, with mean Ψ value as dependent variable. Our strategy was to fit a simple model first, and compare it to increas-

ingly complex models, using a log likelihood test for goodness of fit. Table 3 summarises models and indicates whether they are different from the simplest one (Model 0).[2]

Model	Fixed effects	Random effects	Fit	p
0	Uncertainty	item	NA	NA
1	Uncert.	participant	0	1
2	Uncert.	participant + item	0.916	$> .3$
3	Uncert. + Lang.	item	3.31	$> .06$
4	Uncert. + Prop. + Lang.	item	3.98	$> .3$
5	Uncert. $\times$ Prop. + Lang.	item	4.32	$> .2$
6	Uncert. $\times$ Prop. $\times$ Lang.	item	5.45	$> .4$

Table 3: Linear mixed effects models. Goodness of fit tests compare models to Model 0 using -2 log likelihood

Model 0 is a simple model incorporating only Uncertainty as fixed effect, with item as random effect. This was found to have a high goodness of fit relative to a model with only the intercept and no effects (log $\lambda = 152.4$). The linear mixed effects analysis for this model showed a strong main effect of Uncertainty on Ψ values ($t = 4.887, p < .001$). No subsequent model provided a better fit: Model 1, which incorporates participant as the only random effect, and Model 2, which incorporates both item and participant, are no better, suggesting that the variance among participants was marginal, unlike that of items (scenarios). The impact of different scenarios is likely due to the difference between those where event times were dates and those using clock times – the former are inherently 'fuzzier' since they involve a larger temporal interval.

Once item was established as the only significant random effect, we tested several other models in-

---

[2]The χ^2 values in the table are the $-2LL$ values.

	No uncertainty				1 uncertain proposition				2 uncertain propositions		
	after	before	during	simple	after	before	during	simple	after	before	during
en	0.543 (0.48)	0.672 (0.44)	0.505 (0.50)	0.706 (0.43)	0.217 (0.34)	0.340 (0.43)	0.594 (0.41)	0.159 (0.34)	0.183 (0.36)	0.106 (0.28)	0.604 (0.47)
fr	0.554 (0.48)	0.672 (0.44)	0.411 (0.49)	0.728 (0.41)	0.375 (0.39)	0.311 (0.45)	0.562 (0.44)	0.051 (0.17)	0.185 (0.38)	0.315 (0.44)	0.502 (0.47)
mt	0.736 (0.36)	0.743 (0.38)	0.492 (0.47)	0.656 (0.42)	0.308 (0.38)	0.346 (0.46)	0.434 (0.46)	0.186 (0.33)	0.360 (0.48)	0.255 (0.42)	0.454 (0.47)
overall	0.614 (0.43)	0.696 (0.41)	0.471 (0.48)	0.697 (0.42)	0.298 (0.37)	0.332 (0.43)	0.530 (0.43)	0.1325 (0.30)	0.241 (0.41)	0.2225 (0.39)	0.522 (0.46)

Table 2: Mean Ψ values across languages and conditions (standard deviation in parentheses)

corporating more fixed effect combinations. The main effect of Uncertainty persisted, but Model 3 found only a marginal main effect of language ($t = 1.818, p = .06$) and Model 4 showed no main effect of Proposition Type ($t = 0.811, p > .4$). None of the interactions (Models 5 and 6) yielded a better fit. This replicates the finding of Portet and Gatt (2010), who also found no effect of Proposition Type and no interactions. Perhaps more strikingly, there was no significant difference among participants across the three different languages, suggesting that suggesting that, in our data, the language used to describe scenarios didn't affect uncertainty judgements much. Note that this does not imply that linguistic expressions across languages do not differ, only that for a given set of facts associated with a scenario, the level of subjective uncertainty was independent of the language in which that scenario was described.

	r	p
fr	.45	$< .001$
en	.55	$< .001$
mt	.42	$< .001$
overall	.62	$< .001$

Table 4: Correlations between computed and elicited Ψ judgements.

This finding is encouraging, as it suggests that, to the extent that the reasoning formalism described in Section 3 adequately matches human judgements, it can be used to compute possibility and necessity values (though not their mapping to expressions) independently of the target language in which a given scenario is described. To test this, we computed the Π and N values for each scenario using the reasoning engine described in Section 3, making two assumptions: (i) if a scenario stated that an event occurred at a specific time (or within a fuzzy interval), the event was represented with that time or interval as its start time; (ii) we assumed that, over a given fuzzy interval, the possibility distribution for an event was uniform, that is, if an event was stip-

ulated as having started between t_0 and t_1, it was equally possible/necessary during any subinterval of $[t_0, t_1]$. From the computed values for Π and N the value of Ψ was derived and correlated to the mean Ψ value obtained from participants. Table 4 summarises the correlations for each language, and overall. All correlations were positive and highly significant, and higher when averaged over all languages. The value of $r = .62$ for the 'overall' correlation suggests that we can account for approximately ($.62^2 =$) roughly 40% of the variance in the data. While this is not perfect, it does suggest that the model is on the right track.

4.1.2 Choice of linguistic expression

To address our second question, we attempted to predict the choice of expression made by participants from their subjective uncertainty ratings. This was done for each language separately. Means and frequencies are displayed in Table 5.

In all three languages, the table suggests a clustering of expressions, with higher Π and N for the default, *must* and *definitely* cases, and lower values for *may* and *possibly*. However, there are also divergences: in French, the counterpart for *definitely* has a much lower N than in English or Maltese. French *may* and *possibly* also have lower Π values. Maltese Π values for *may* and *possibly* are also closer to those for other expressions than they are in French or English, although the corresponding N values are similar.

Since our aim is ultimately to develop a function that can map from a particular level of subjective uncertainty to a modal expression in a given language, we modelled these results using a multinomial logistic regression (essentially, a Maximum Entropy model). This amounts to treating our problem as a classification problem: given a scenario and a temporal relation, with associated Π and N values, what linguistic expression do these values map to? Our model used the *default* as the reference category, to which others are compared. We simplified the

English			French			Maltese		
	Π	N		Π	N		Π	N
default (27)	0.96	0.82	default (50)	0.97	0.92	default (39)	1	0.76
must (20)	1	0.92	doit (10)	1	0.81	bilfors li (18)	0.94	0.82
definitely (27)	1	0.94	sûrement (19)	0.96	0.49	żgur (26)	0.98	0.78
may (54)	0.86	0.09	pourrait (38)	0.71	0.09	jista' jkun li (55)	0.95	0.14
possibly (28)	0.90	0.11	peut-être (35)	0.73	0.04	forsi (20)	0.90	0.16
negation (23)	0.58	0.04	negation (17)	0.24	0	negation (12)	0.29	0

Table 5: Mean Π and N values by phrase choice. Frequency of each choice is in parentheses.

modelling process by dividing the subjective Π and N ratings into four intervals at increments of 0.25 (i.e. the new coding grouped together $\Pi < .025$, $0.25 \geq \Pi < 0.5$ etc), effectively recoding the predictor variables into categorical ones.

For both English and French, the model incorporating both Π and N yielded an excellent goodness of fit (English: model $\chi^2 = 265.03, p < .001$; French: $\chi^2 = 205.46, p < .001$). However, this was not the case for Maltese, where the combined model was not significantly better than a model containing only the intercept. For this language, a model with only N as predictor turned out to be better ($\chi^2 = 134.87, p < .001$). This is relatively unsurprising, considering that the possibility values for the Maltese data are quite consistently high, with the exception of the negated expressions. This may reflect a genuine difference between Maltese and the other two languages under consideration; however, given that the samples used in the present study were relatively small, further testing will be required to establish the reliability of this finding.

4.1.3 Lexical choice of modals in NLG

A regression model such as the one developed above can be used to classify particular instances (combinations of Π and N values), to identify the best modal expression to use to express the temporal uncertainty. To take an example suppose the reasoning engine predicts $\Pi = 1$ and $N = 0$ for the proposition *the police were on the scene before the robbers left the bank*. The model for English predicts no change in the likelihood of choosing the default expression (i.e. the unmodalised proposition) where possibility values are high, all other things being equal. However, in the present case, the low necessity value substantially decreases the odds associated with the default. In this case, therefore,

the model would swing the choice in favour of that expression whose probability increases, relative to the default, as necessity decreases. In this case, the most likely such expression is *possibly*. The model would work in the same way for the other two languages under consideration. Furthermore, given that our results suggest that the actual uncertainty ratings for scenarios are independent of language (Section 4.1.1), we hypothesise that extending the model to other languages would not require substantial alterations to the reasoning formalism described in Section 3, but only to the specific classification model.

5 Conclusions

This paper presented a formalism to reason with temporal uncertainty and a model to map from uncertainty to modal expressions in different languages. Our data shows that subjective uncertainty varies as a function of the temporal uncertainty associated with events in a scenario; moreover, subjective uncertainty correlates well with the values computed by our model. Although we find no evidence of a strong effect of participant variation in our data, in future work we plan to investigate to what extent subjective uncertainty differs between participants using larger samples, as previous work has shown that individual reasoning strategies may differ (Benferhat et al., 2005).

We also described a logistic regression model to predict the best expression in a particular language given a specific degree of subjective uncertainty. The experimental data suggests that there are substantial differences between the sets of expressions tested for the three languages. More data from more participants will be required to validate it and this is our aim in the medium term, in addition to extending our model to cover more linguistic expressions.

Acknowledgments

References

J. F. Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

Alexander Artikis, Georgios Paliouras, Franois Portet, and Anastasios Skarlatidis. 2010. Logic-based representation, reasoning and machine learning for event recognition. In *DEBS 2010*, pages 282–293.

S. Barro, R. Marín, J. Mira, and A. Paton. 1994. A model and a language for the fuzzy representation and handling of time. *Fuzzy Sets and Systems*, 61:153–175.

S. Benferhat, J.F. Bonnefon, and R. da Silva Neves. 2005. An overview of possibilistic handling of default reasoning, with experimental studies. *Synthese*, 146:53–70.

M. Campos, A. Cárceles, J. Palma, and R. Marín. 2002. A general purporse fuzzy temporal information management engine. In *EurAsia-ICT 2002*, pages 93–97.

R. Dechter, I. Meiri, and J. Pearl. 1991. Temporal constraint networks. *Artificial Intelligence*, 49:61–95.

D. Dubois, H.A. Allel, and H. Prade. 2003. Fuzziness and uncertainty in temporal reasoning. *Journal of Universal Computer Science*, 9(9):1168–1194.

M. Elhadad. 1995. Using argumentation in text generation. *Journal of Pragmatics*, 24:189–220.

N. Evans and S.C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32:429–492.

F. Gao, S. Sripada, J. Hunter, and F. Portet. 2009. Using temporal constraints to integrate signal analysis and domain knowledge in medical event detection. In *12th Conference on Artificial Intelligence in Medicine (AIME 09)*, volume 5651 of *LNAI*, pages 46–55.

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

P. Johnson-Laird. 1978. The meaning of modality. *Cognitive Science*, 2:17–26.

L. Karttunen. 1972. Possible and must. In J. Kimball, editor, *Syntax and Semantics*, volume 1. Academic, New Yowk.

R. Klabunde. 2005. When must should be chosen. In *Proceedings of the 10th European Workshop on Natural Language Generation (ENLG-05)*.

R. Klabunde. 2007. Lexical choice for modal expressions. In *Proceedings of the 11th European Workshop on Natural Language Generation (ENLG-07)*.

A. Kratzer. 1977. What *must* and *can* must and can mean. *Linguistics and Philosophy*, 1:337–355.

A. Kratzer. 1981. The notional category of modality. In H-J Eikmeyer and H. Rieser, editors, *Words, Worlds and Contexts: New Approaches to Word Semantics*. Walter de Gruyter and Co., Berlin.

J. Palma, J.M. Juareza, M. Camposa, and R. Marina. 2006. Fuzzy theory approach for temporal model-based diagnosis: An application to medical domains. *Artificial Intelligence in Medicine*, 38(2):197–218.

F.R. Palmer. 2001. *Mood and modality*. Cambridge University Press, Cambridge, 2 edition.

A. Papafragou. 1998. Inference and word meaning: The case of modal auxiliaries. *Lingua*, 105:1–47.

A. Papafragou. 2006. Epistemic modality and truth conditions. *Lingua*, 116:1688–1702.

F. Portet and A. Gatt. 2010. Towards a possibility-theoretic approach to uncertainty in medical data interpretation for text generation. In D. Riano, A. ten Teije, S. Miksch, and M. Peleg, editors, *Knowledge Representation for Health-Care: Data, Processes and Guidelines*, LNAI 5943. Springer, Berlin and Heidelberg.

F. Portet, E. Reiter, A Gatt, J. Hunter, S. Sripada, Y. Freer, and C. Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*.

P. Portner. 2009. *Modality*. Oxford University Press, Oxford.

E. Raufaste, R. Da Silva Neves, and C. Mariné. 2003. Testing the descriptive validity of possibility theory in human judgments of uncertainty. *Artificial Intelligence*, 148(1-2):197–218.

E. Reiter, R. Robertson, and L. Osman. 2003. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144:41–58.

E. E. Sweetser. 1990. *From etymology to pragmatics*. Cambridge University Press, Cambridge.

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

M. Vanhove, C. Miller, and D. Caubet. 2009. The grammaticalisation of modal auxiliaries in maltese and arabic vernaculars of the mediterranean area. In B. Hansen and F. de Haan, editors, *Modals in the languages of Europe*. Mouton, Berlin.

L. Vila and L. Godo. 1994. On fuzzy temporal constraint networks. *Mathware and soft computing*, 3:315–334.

M. Vilain, H. Kautz, and P. van Beek. 1987. Constraint propagation algorithms for temporal reasoning: A revised report.

L.A. Zadeh. 1965. Fuzzy sets. *Information and Control*, 8(3):338–353.

L. Zhou and G. Hripcsaka. 2007. Temporal reasoning with medical data — a review with emphasis on medical natural language processing. *Journal of Biomedical Informatics*, 40(2):183–202, April.