Managing Uncertainty in Time Expressions for Virtual Assistants

Xin Rong[†*], **Adam Fourney**[‡], **Robin N. Brewer**[¶*], **Meredith Ringel Morris**[‡], **Paul N. Bennett**[‡]
[†]University of Michigan, [‡]Microsoft Research, [¶]Northwestern University
ronxin@umich.edu, adamfo@microsoft.com, rnbrewer@u.northwestern.edu, merrie@microsoft.com, pauben@microsoft.com

ABSTRACT

"Remind me to get milk *later this afternoon.*" In communications and planning, people often express uncertainty about time using imprecise temporal expressions (ITEs). Unfortunately, modern virtual assistants often lack system support to capture the intents behind these expressions. This can result in unnatural interactions and undesirable interruptions (e.g., having a work reminder delivered at 12pm when out at lunch, because the user said "this afternoon"). In this paper we explore existing practices, expectations, and preferences surrounding the use of ITEs. Our mixed methods approach employs surveys, interviews, and an analysis of a large corpus of written communications. We find that people frequently use a diverse set of ITEs in both communication and planning. These uses reflect a variety of motivations, such as conveying uncertainty or task priority. In addition, we find that people have a variety of expectations about time input and management when interacting with virtual assistants. We conclude with design implications for future virtual assistants.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords

Time expressions; virtual assistants; uncertainty management

INTRODUCTION

Virtual assistants (e.g., Apple's Siri, Microsoft's Cortana, Amazon's Alexa, or Google Now) allow people to use natural language to access a device's commands, settings, and integrated services. In many cases, these interactions require that the user communicate a date and/or time. Examples of such scenarios include: creating appointments, setting up reminders, or asking virtual assistants about weather, news, sports, etc. Whether issuing commands or retrieving information, users are encouraged, by convention and instruction [18, 37], to structure their utterances as if they were talking with a real person. However, as we show, these types of interpersonal communications often give rise to temporal expressions that are imprecise, nuanced, and ambiguous. To adopt this style in a virtual assistant is to inherit the challenges of recognizing and managing temporal uncertainty.

Type	Expression
Imprecise	a little more than a week later tomorrow evening this weekend
Precise Date	21 September 2016 tomorrow this Thanksgiving
Precise Time	21 September 2016, 8:00pm EDT 8:00 a.m. tomorrow in 15 minutes

Table 1. Examples of imprecise and precise time expressions: a precise date resolves to the entirety of a specific calendar day when taken *in conjunction with a reference time* (e.g., the time the expression was uttered). A precise time resolves to a date and time down to the minute.

In contrast, modern virtual assistants often insist that dates and times be specified precisely, or will rigidly map a limited set of temporal expressions to predetermined wall-clock times (e.g., mapping "morning" to 7:00am). While early resolution of these expressions may be the easiest solution for system designers and implementers, this strategy can lead to breakdowns in the user experience. First, we show that people are often strategic in their use of temporal expressions so as to convey their own uncertainty, commitment, or task priority. In these scenarios, insisting that people input specific times or dates may increase the burden of using the system. Likewise, overly literal interpretations of temporal expressions may result in reminders or notifications being delivered at inappropriate times (e.g., a user specifying "this afternoon" may not want or expect a reminder to be delivered at precisely 12:00pm).

In this paper we investigate these and related issues pertaining to imprecise temporal expressions (ITEs) in interactions with virtual assistants. For the purpose of this work, we define an ITE as a temporal expression that neither resolves to a *precise date* (i.e., the entirety of a calendar day), nor a *precise time* (i.e., a time with both hour and minute specified), as outlined in Table 1. Our research on ITEs is structured around answering the following research questions (see Figure 1):

- **Motivations**: When do people prefer to use imprecise time expressions? What motivations underlie these choices?
- **Manifestations**: How are temporal expressions manifest in interpersonal communications? What can these occurrences

[*]Work done while at Microsoft Research.

CHI 2017, May 6-11, 2017, Denver, CO, USA.

http://dx.doi.org/10.1145/3025453.3025674

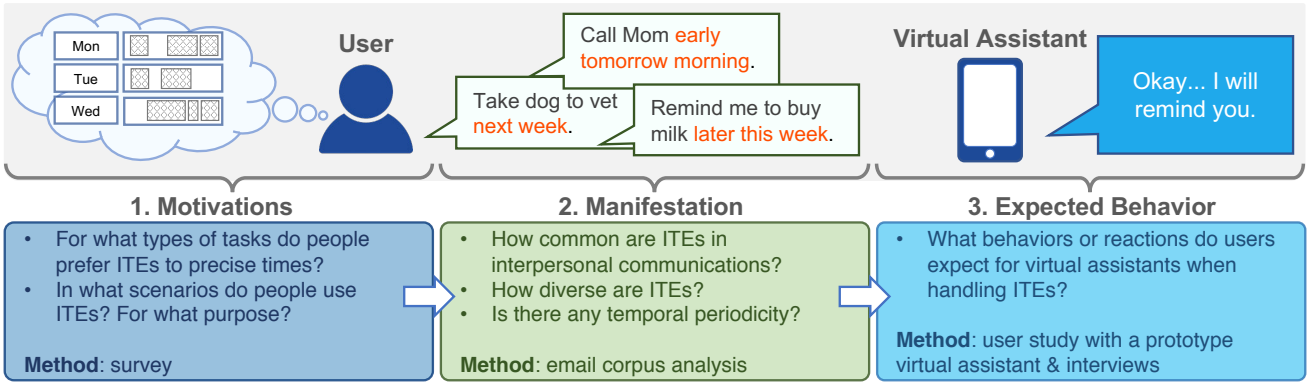

Figure 1. Three aspects of ITE management: user motivations, ITE manifestation, and the expectations for virtual assistants.

tell us about the prevalence and variety of expressions a virtual assistant might encounter in the wild?

- **Expectations**: What are people's expectations when using temporal expressions with virtual assistants?

Given the varied nature of these research questions, our investigations employ a mix of methods. We begin by reporting the results of a survey that broadly characterizes people's motivations for using precise or imprecise temporal expressions when interacting with virtual assistants. Next, given that interactions with modern virtual assistants are styled around natural conversations, we report how temporal expressions are manifested in a large corpus of interpersonal email communications. Finally, we describe a prototype virtual assistant/productivity application we developed to serve as a design probe for this research. This assistant provides support for the broad set of expressions and interpretations obtained from our survey and email studies. We describe the themes that emerged from interviews with 14 participants who used this interface. These in-person interviews were designed to gather a deeper, and more nuanced view of peoples' expectations when using temporal expressions in this context.

Taken together, results from these studies generate insights about the sources of uncertainty that arise with temporal expressions, and describe why management of these temporal uncertainties is important for the development of virtual assistants. In the remainder of this paper we review related work, then introduce the three studies outlined above. We present a set of design implications, then more broadly discuss the implications and limitations of this research.

RELATED WORK

We review related work that handles detection and reasoning about imprecise time expressions (ITEs), as well as research on how virtual assistants manage prospective tasks and handle uncertainties in user input.

Extracting and Reasoning About Time Expressions

Existing work has recognized that ITEs occur in a variety of contexts (e.g. news corpora, electronic health records) [3, 9, 31, 36]. Multiple strategies and solutions have been proposed to both detect and reason about these entities. For detecting temporal expressions, rule-based systems have been shown to perform well, including GUTime [19], HeidelTime [33], SUTime [6], and HINX [36], which can annotate the extracted time phrases as standard markups, such as TIDEs [10] or TimeML [26]. Probabilistic frameworks are an alternative, and have the advantage of leveraging contextual information, such as verb tenses in the sentence, to determine proper temporal interpretation [1, 16]. Tissot [36] combines fuzzy set theory and crowdsourcing to obtain normalization for time phrases that are grounded by human judges. We do not focus on improving the extraction of imprecise time expressions – we mainly rely on existing systems (SUTime) to do such work, but also show how ITEs can be expanded using an existing corpus. Instead, our primary focus is on studying the user intent behind embedding uncertainty in time expressions.

For reasoning about ITEs, solutions have been proposed to place events in temporal sequences to help answer queries related to historical events [9, 29], or to solve constraint-based scheduling problems such as taxi dispatching [34]. These works leverage fuzzy set theory [40, 41] and the fuzzy temporal interval relational models [30] to handle the impreciseness of time representations. However, existing work has not explored the original motivations for people's use of imprecise times, nor has it explored their use in interactive settings.

Time Management by Virtual Assistants

Past work has also explored how virtual assistants can manage time schedules in personal or team settings [7, 8, 14, 15, 28]. In particular, past work has explored triggering reminders based on contextual signals, such as location [17, 32], activities [7], and time [12]. Studies have also identified strong patterns in people's time selections when using reminder apps to manage routine activities [12, 32]. For example, a large-scale log analysis of a commercial virtual assistant has shown that a reminder's creation time and its content are predictive of the user-specified notification time [12].

To automate people's routine activity management, intelligent assistants have been built to execute routine tasks, including handling emails [11], scheduling meetings [39], and managing calendars [2, 4, 22, 28]. The framework of such systems can be built with a Belief-Desire-Intention (BDI) model [27, 35, 38].

User preference modeling, online learning, and constraint reasoning constitute important components [4]. However, these past works have relied on people using precise times. In this paper, we have a special focus on the use of ITEs and we reveal that strong patterns and preferences exist here as well.

Uncertainty Management by Virtual Assistants

Another set of work addresses uncertainty in conversations between humans and intelligent agents [24, 25]. For example, Paek and Horvitz [24] propose a multi-layer computational framework consisting of Bayesian networks for guiding actions given conversational inputs with uncertainty. Pineau et al. [25] propose a mobile robotic assistant system that employs a Partially Observable Markov Decision Process to handle uncertainty in observations. Our work can help to motivate and ground such efforts in data obtained from multiple sources reflecting users' actual usage, preferences, and opinions about time scheduling with uncertainty.

Closest to our work is research done by Martin and Holtzman [20], who propose Kairoscope, a time management approach that is based on relative event sequences instead of precise times. It allows the user to specify imprecise times such as "tomorrow" or "later this week" and uses peer-to-peer agent communication to negotiate and dynamically reschedule event times. Our work compliments their research by greatly expanding the set of expressions to consider, and by exploring the reasons and motivations for using ITEs in the first place.

In summary, our contributions beyond the existing work include: (1) new data analysis that reveals the motivation behind the usage of ITEs; (2) characterizing the prevalence of ITEs in computer-mediated conversations by measuring the frequency, variety, and periodicity of a large set of ITEs; (3) understanding people's expectations of the behaviors of virtual assistants in situations where ITEs are used; and (4) design implications specific to managing uncertainties for modern virtual assistant systems based on the gathered data and observations.

MOTIVATIONS FOR USING IMPRECISE TIME

In this section, we address the first research question by reporting the results of a large survey. The survey sought to characterize the factors that underlie people's choices of temporal expressions in interactions with virtual assistants.

Procedure

The survey consisted of two sections, and was emailed to a random set of 6,000 employees within a large IT corporation. Demographic information was collected at the end of the questionnaire. We describe the survey's components below.

Levels of temporal precision and certainty

The first section presented respondents with concrete situations that required them to consider how they would specify times and dates to a virtual assistant. This section served the dual role of grounding the main content questions of the survey, while also gathering data about the types of scenarios most likely to evoke temporal expressions at varying levels of precision. The following prompt was presented:

> Imagine that, at this particular moment, you are about to use a virtual assistant to create a to-do list. Further suppose that, upon adding each item to the list, the virtual assistant asks you to specify a date and/or time. For each to-do item, please complete the following sentence by selecting from the drop-down menus below:
>
> "I am most comfortable answering with a date / time that is accurate to within..."

Possible sentence completions are listed in Figure 2 (horizontal axis), and range from as low as a minute to as high as a year.

Respondents were then presented with a list of five to-do items, and five blank spaces in which they could input their own items. Respondents were asked to input at least two of their own items. Pre-populated items were randomly selected from the five most common reminder categories identified by Graus et al. in [12]. The tasks (and their categories) were: pick up laundry (run errand), pay phone bill (chores), call mom (communicate), start cooking (manage process), and take medicine (eat/consume).

When and why ITEs are preferred

The second section of the survey was designed to more generally explore the situations that might trigger people to use imprecise temporal expressions. Respondents were asked to more generally reflect on the situations where they might use imprecise temporal expressions over giving precise times. The following prompt was presented:

> Modern virtual assistants can help you manage your to-do lists by associating precise times with each item. Like an alarm clock, these precise times specify both hour and minute (e.g., 8:05am). Mirroring human assistants, virtual assistants can also increasingly cope with imprecise time expressions (e.g., "in a few hours", "this evening", "later this week", "on the weekend", etc.)
>
> Rate your agreement to each of the following statements.
>
> "I prefer to use imprecise time expressions (e.g., 'in a few hours', 'this evening', etc.) when ..."

We then asked the respondents to respond to 11 sentence completions (abbreviated in Figure 4). Two additional options served as catch-all categories: "*for most tasks/todo-items*" or "*in few, or almost no, situations*." All items were scored on a 5-point Likert scale (strongly disagree = 1, strongly agree = 5).

Results

338 respondents completed the entire survey, while an additional 182 participants provided partial responses (response rate: 8.0%, completion rate: 65.0%). In our analysis, we only consider those 338 who completed the survey in its entirety.

73.7% (n = 249) of the respondents were male, and 83.4% (n = 282) reported having a bachelor's or advanced academic degree. Respondents reported occupying a diverse set of roles within the company: 24.6% (n = 83) were software developers, 21.6% (n = 73) worked in sales or marketing, and 16.9% (n = 57) worked in managerial positions. The remaining 37.0% of respondents (n = 125) worked in a number of specific roles,

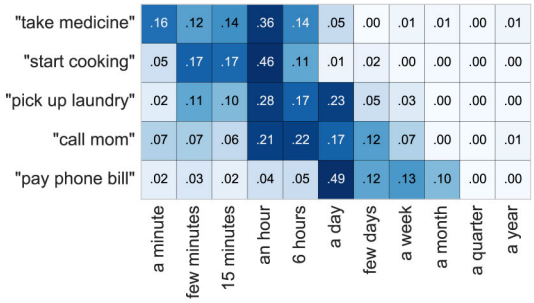

Figure 2. Respondent preference for various levels of temporal precision over the five sample to-do items provided in the questionnaire.

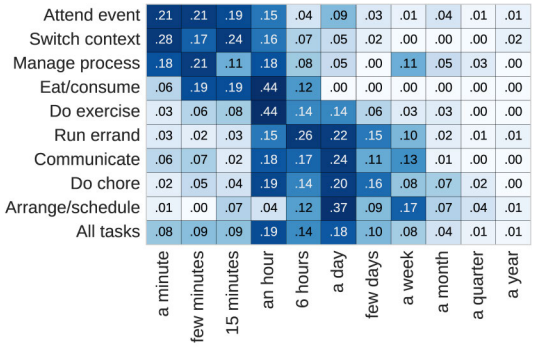

Figure 3. Respondent preference for various levels of temporal precision over the 865 respondent-provided to-do items. Items are organized into nine task categories, adapted from [12].

including: technical writers, supply chain engineers, security consultants, legal advisers, etc.

Regarding prior experience with virtual assistants, 17.8% (n = 60) of respondents indicated that they interacted with a virtual assistant (any one of Siri, Google Now, Cortana, and Echo/Alexa) earlier the same day; 30.8% (n = 104) earlier in the week; 15.1% (n = 51) earlier in the month; 22.5% (n = 76) more than a month ago; and 13.9% (n = 47) never. Overall a majority of the respondents (63.5%) had interacted with a virtual assistant within the past month.

Levels of temporal precision and certainty

Figure 2 summarizes responses for the pre-populated to-do items. Figure 3 summarizes responses for the 865 user-provided items, manually classified into nine task categories adapted from [12].[1] Together, these figures express the two main findings from this portion of the questionnaire: (1) temporal uncertainty varies considerably by task type, and (2) respondents were rarely comfortable expressing minute-level precisions (occurring in only 2% to 16% of responses for the pre-populated items listed in Figure 2, and in 9% of the user-provided items in Figure 3). Recognizing the possibility that some respondents may have conflated minute-level responses with hour-level responses (e.g., for items like "8:00am"), we further report that hour and minute-level precisions combine to account for at most 52% of responses.

[1]The open-ended survey data were manually labeled by one of the authors.

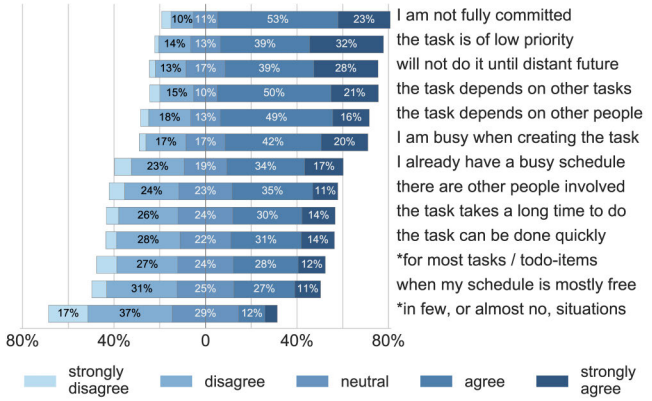

Figure 4. List of situations where people prefer to use imprecise time expressions over precise time expressions. Items are scored on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). Starred (*) items are two catch-all categories.

Further analysis of user-provided tasks reveals additional nuance. For example, respondents were more comfortable with higher levels of precision for tasks involving the attendance of an event (e.g., "meet at restaurant") than when scheduling or arranging events (e.g., "make reservations for dinner"). This difference is highly statistically significant (Mann–Whitney U = 8618.5, two-tailed $p < 0.001$). Likewise, we find that respondents preferred to be less precise when referring to chores (e.g., "*organize clothes*") or running errands (e.g., "*go to the post office*"). These findings appeal to our intuitions about *external dependencies* (scheduling), and *task priority* (chores and errands), which we explore further in the next section.

When and why ITEs are preferred

Notably, we find that 40.2% (n = 136) of respondents reported that "*(they) prefer to use imprecise temporal expressions for most tasks / to-do items.*" This is significantly more than the 16.9% (n = 57) of respondents who reported that "*(they) prefer to use imprecise temporal expressions in few, or almost no, situations*" (Mann–Whitney U=67656.0, $p < 0.001$).

Further investigation reveals that the leading situation where people prefer to use imprecise time is when they are not fully committed to the task (Figure 4). Five other situations closely follow: "*when the task is low priority*", "*when the task will not be executed until the distant future*", "*when the task depends on other tasks*", "*when the task depends on other people*", and "*when I am busy at the time of creating the task*".

Finally, we studied the responses of the 57 respondents who felt that imprecise temporal expressions were preferred in "*almost no situations*." Even within this group, respondents were more likely to prefer imprecise temporal expressions when: "*(they) have not fully committed to the task*" (75% agreement), "*when the task will not be executed until the distant future*" (74% agreement), and "*when the task is low priority*" (65% agreement). In each of these three cases, a Binomial test finds these preferences to be statistically significant at the 0.05 level ($p < 0.001$, $p < 0.001$, and $p = 0.033$, respectively).

Discussion

From the results of the survey's first section, we conclude that when managing common tasks, people are only occasionally comfortable with minute-level precision. At present, however, minute-level precision is required by contemporary virtual assistants when scheduling notifications or reminders [12].

Responses to the second section of the survey mirrored those from the first section, reinforcing the need for virtual assistants to consider task priority and external dependencies when interpreting ITEs. To this end, we note that the situation "*when there are other people involved*" is similar to "*when the task depends on other people*", but lacks the language suggesting a dependency. Accordingly, we find that the former expression has much lower preference for imprecise temporal expressions than the latter (p<0.001, by Mann–Whitney U). Additionally, the tendency to prefer imprecision "when busy" at task creation time suggests that cognitive load may also be a factor.

In summary, we find that there are many common scenarios in which people are likely to prefer imprecise time expressions when interacting with virtual assistants. Imprecision was found to be most appropriate for tasks of low commitment or priority or when tasks involved external dependencies. Next, we characterize how these ITEs are likely to manifest.

MANIFESTATION OF IMPRECISE TIME

To answer our next research question, we analyzed a large email corpus consisting of interpersonal communications. By extracting and analyzing the characteristics of temporal expressions found in this corpus, we can devise a set of ITEs to consider in the later stage of the research and also draw design implications for virtual assistants.

Data

Our analysis used the publicly available Avocado dataset, a corpus of emails exchanged between 279 correspondents "of a defunct information technology company" [23]. Although the language used in email conversations may differ from that with virtual assistants, the Avocado corpus exhibits a number of desirable properties. In particular, as mentioned above, because the email exchanges occurred between human correspondents, the corpus affords an opportunity to observe how temporal expressions occur without user expectations of the limitations of existing virtual assistants. Moreover, since the email conversations contain temporal metadata (i.e., sent time stamps), we are able to analyze the temporal patterns of the occurrences of any given expression.

Procedure

Preprocessing

The dataset required several cleaning and normalization steps prior to linguistic processing. First, we extracted the text in the email body of each email, discarding the subject, senders, recipients, and other header fields. We then removed quoted content from email replies (or forwarded messages) so that each original message was analyzed only once. This resulted in 379,332 valid emails. We then also removed 74,053 messages that appeared to be spam or machine-generated. 305,279 emails remained after all the filtering steps. Finally, all text content was transformed to lower case and then tokenized for further analyses.

Extracting temporal expressions

Our analyses required that we extract a broad range of temporal expressions from the Avocado corpus. To this end we used SUTime, a rule-based temporal expression tagger [6] capable of detecting both precise dates or times (e.g., "October 3rd"), and a limited set of ITEs (e.g., "Saturday morning").

To expand the coverage to a broader set of temporal expressions, we performed an additional amplification procedure. First, we identified common keywords, prefixes, and suffixes occurring in the temporal expressions already identified by SUTime. We then computed the cross product over these three sets, thus generating additional candidate expressions. Finally, we again filtered the results on the Avocado corpus, keeping only those expressions occurring at least once.

Categorizing temporal expressions

Having extracted temporal expressions from the Avocado dataset, the second step of the procedure was to classify each expression as either cleanly resolving to a precise date/time, or as an expression that was unresolvable (i.e., an ITE). A rule-based classifier was developed for this purpose, and operated as follows – temporal expressions were classified as precise if they met *all* of the following conditions: (1) identified by SUTime; (2) did not occur near approximation keywords (e.g., "around", or "in about"); and (3) were not ascribed any special annotation by SUTime. For example, SUTime outputs special annotations such as "2017-SU" when parsing the phrase "next summer", and "2016-09-21EV" when parsing the phrase "tomorrow evening". Temporal expressions failing any of these checks were labeled as ITEs. In the remainder of this section, we are primarily concerned about ITEs.

Human interpretation of temporal expressions

Finally, while the Avocado corpus contains a record of when imprecise temporal expressions were invoked (email *sent* time), it lacks information about how those expressions may have been intended or interpreted. To gather these mappings, we tasked crowd workers with labeling a small set of 53 extracted expressions. A sample labeling task reads: "*assuming that today is Monday Jan 11th, what does 'later this week' mean to you?*" Workers were then asked to choose a date range by selecting a start date and an end date from a calendar widget. For each expression considered, a human intelligence task (HIT) was prepared for all seven days of the week. For each question, 50 judgments were collected. In total, 308 workers contributed to judgments. All workers resided in English-speaking countries, and were recruited from the Clickworker.com crowd-sourcing platform.

Results

ITEs are common

Chief among our findings is the observation that 36.2% of all temporal expressions in the email corpus could not be automatically resolved to specific dates or times.

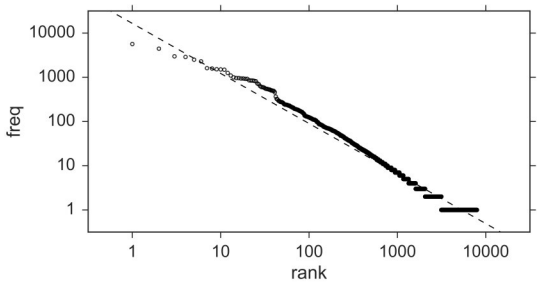

Figure 5. Occurrence frequencies of imprecise time expressions in the Avocado email dataset follow Zipf's law.

Expression	Count
next week	5653
this week	4436
last week	2979
lunch	2881
this morning	2501
a little more than a week later	1
no later than tomorrow noon	1
around the end of this week or early next week	1
morning Friday, or at latest, very early next week	1
lunch tomorrow, or later in the day	1

Table 2. Examples of imprecise temporal expressions. Top: most frequent expressions. Bottom: selected complex expressions.

ITEs follow a power law distribution

In addition to reporting aggregate frequencies (above), we also investigated the occurrence rates of individual temporal expressions. Our analysis reveals that the frequencies of ITEs follow a power-law distribution, with the most common expressions occurring exponentially more often than the least common expressions (Figure 5). Importantly, 79.7% of all expressions are observed 3 or fewer times, and 60.4% of all expressions are observed only once. Most expressions occur with low frequency, and many are complex (Table 2).

ITEs exhibit temporal periodicity

Finally, our analysis reveals that a number of expressions present periodicity in their use. To illustrate this pattern, for a given time expression, we look at the sent time of each email containing the time expression. We then count the occurrences of selected popular expressions, binning over the days of week (Figure 6), or over three-hour slots each day (Figure 7). We normalize the counts based on the total number of emails sent on each day, or time block, respectively. For example, Figure 6(a) indicates that the mentions of "later this week" occur most often on Mondays and Tuesdays, and mentions of "early next week" occur most often on Fridays.

Figure 6(b) provides a sanity check – the peak use of a given day of the week (e.g., "Thursday") does not occur on that particular day, or even adjacent days. Note that, on a given Thursday, the word "today" becomes available for some intended uses. Likewise, on Wednesday, the word "tomorrow" becomes available. This explains why the mentions of "Thursday" peak on Tuesday, and the mentions of the other days of week share a similar trend. Figure 6(c) illustrates the distribution of mentions for a set of ambiguous terms. The mentions

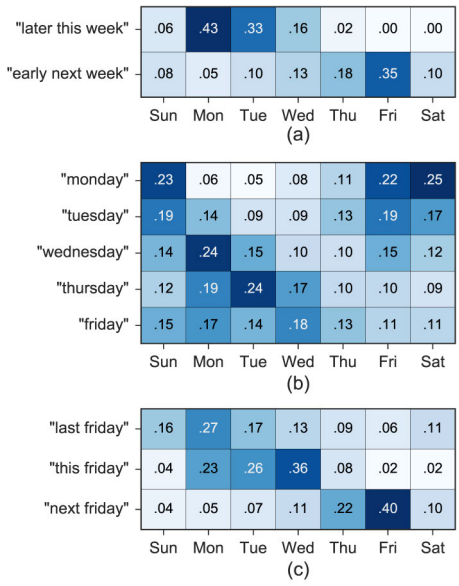

Figure 6. Normalized usage frequency of imprecise time expressions per day-of-week in the Avocado data set.

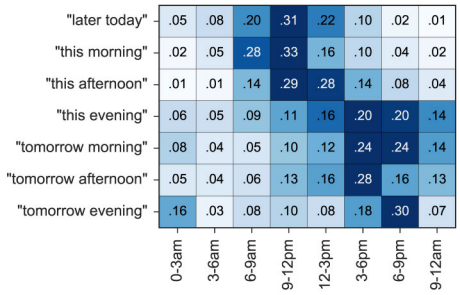

Figure 7. Normalized usage frequency of ITEs per 3-hour slot of a day in the Avocado data set.

of "last Friday" occur most often on Mondays, mentions of "this Friday" occur most often on Wednesdays, and mentions of "next Friday" occur most often on Fridays.

Within the scope of a day, some expressions also exhibit periodicity (Figure 7). For example, the mentions of "later today" occur most often in the morning between 9am–12pm. So do the mentions of "this morning." In comparison, mentions of the parts of day for the next day most often occur in the evenings.

Given the observed periodicities in ITE occurrences, we hypothesized that temporal context may also play a role in people's interpretations of ITEs. To examine this hypothesis, we examined the times ascribed to ITEs by crowd workers. We observed that the interpretation of ITEs varies depending on the time in which it is considered. For example, when the given reference time is early in the work week, the interpretations of the phrase "next weekend" are characterized by a bimodal distribution (Figure 8). Here, some judges interpret the phrase to mean the closest upcoming weekend, while others interpret the phrase to mean the weekend that concludes the following week. However, by Wednesday, much of the ambiguity is gone.

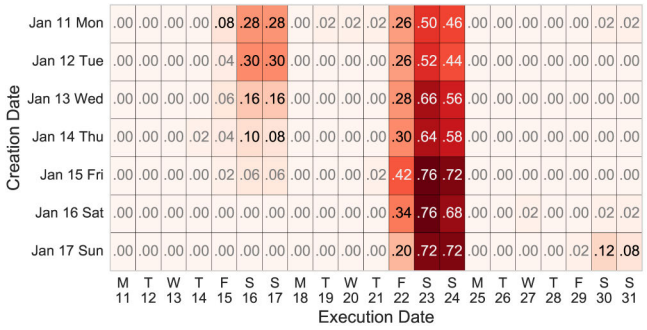

Figure 8. Interpretation of "next weekend" on different days of the week based on the crowdsourced results. The numbers shown are the proportion of workers that include the particular execution date in their range of interpretation.

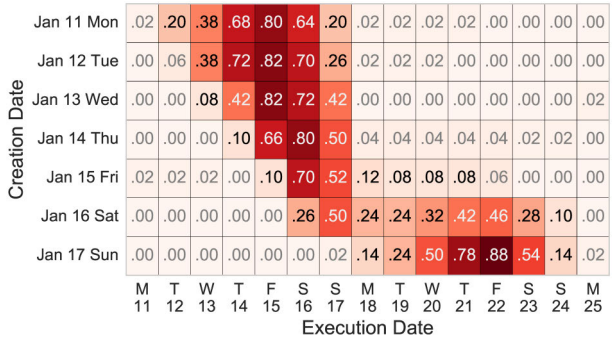

Figure 9. Interpretation of "later this week" on different days of the week based on the crowdsourced results. The numbers shown are the proportion of workers that include the particular execution date in their range of interpretation.

Likewise, Figure 9 summarizes the result we obtained for the interpretation of "later this week." Agreeing with intuition, "later this week" is most often interpreted to mean the remainder of the work week (ending Friday). However, an interesting discrepancy arises on Thursdays. Specifically, from Monday to Wednesday, the interpreted distribution peaks on Friday. However, beginning on Thursday the peak in interpreted times shifts to Saturday. We hypothesize that the crowd workers shifted their estimates because, beginning Thursday, the Friday is most naturally referred to simply as "tomorrow".

Discussion

So far we have presented three main findings from the email corpus analysis. We now discuss the implications of these results.

First, ITEs being very common implies that if virtual assistants are unable to interpret these expressions, they may be missing a third or more of all temporal expressions occurring in the natural communications they seek to understand and emulate.

Second, ITEs being very diverse implies that virtual assistants must be effective in resolving *long-tail* temporal expressions. While heuristics can be developed to resolve the most common ITEs, such as "next week," "this morning," "this afternoon," "this weekend," and "tomorrow morning," inspection of long-tail expressions reveals more complex temporal expressions (Table 2) that may create difficulties not only for extraction, but also for downstream interpretation and action (e.g., triggering reminders).

Third, ITEs exhibiting periodic regularities implies that date and time are a crucial context for correctly interpreting these expressions. Designers of virtual assistant are advised to consider these issues when developing interactions. For example, when the user says "later this week" on a Friday, which is a rare event according to its past occurrence frequencies, the system may verify with the user by asking what she really means. Similarly, if the user says "next weekend" on a Monday, which is very ambiguous according to the interpretation model, the system may also actively verify with the user.

In summary, and in response to our second research question, we find that ITEs represent a large class of temporal expressions occurring in computer-mediated communications, both in terms of frequency and in diversity. Moreover, we find that both the past occurrences and present interpretations of ITEs are themselves governed by temporal processes. These findings have immediate implications for the design of virtual assistants, which we have outlined above, and later revisit.

EXPECTATIONS FROM VIRTUAL ASSISTANTS

To answer our final research question, we conducted an in-person interview study with 14 participants. The study was designed to gather a more nuanced view of peoples' expectations of virtual assistants in scenarios where interactions involve temporal expressions.

Design Probe Apparatus

A prototype virtual assistant system was developed to serve as a design probe, with the goal of grounding the interviews, and guiding participants to speak in concrete terms about their expectations around ITEs. It is the dialog with participants, and not the system itself, which forms a primary contribution of this paper. To this end, the virtual assistant's functionality was limited: the system allowed the user to create and manage a list of memos and reminders; it recognized ITEs using the same method as in the email corpus study; and, it used heuristic rules to map ITEs to actual time intervals (Figure 11); however, it could not deliver reminders as notifications.

Nevertheless, one feature of the prototype distinguishes it from existing virtual assistants. Namely, the prototype prominently displayed a curated list of items deemed relevant to the current moment. This region of the display was referred to as the *Corkboard* (Figure 10), and it allowed multiple items with imprecise and overlapping time periods to be displayed together without evoking the sense of conflict that arises in calendaring applications when two entries overlap. When creating a memo or a to-do item (Figure 11), the user could schedule the time period in which the given item would be displayed in this region by either: (1) including temporal expressions in the text content of the item (Figure 11, e.g., "get coffee after work"); or (2) manually selecting from a list of common expressions (e.g., "today", "tomorrow", etc.); or (3) picking a precise date or time from a calendar or time picker widget.

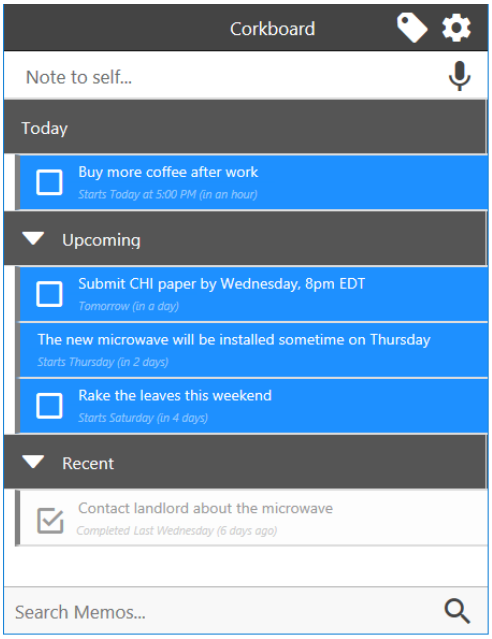

Figure 10. Our virtual assistant's *Corkboard* screen displays a short list of items relevant to a given moment in time.

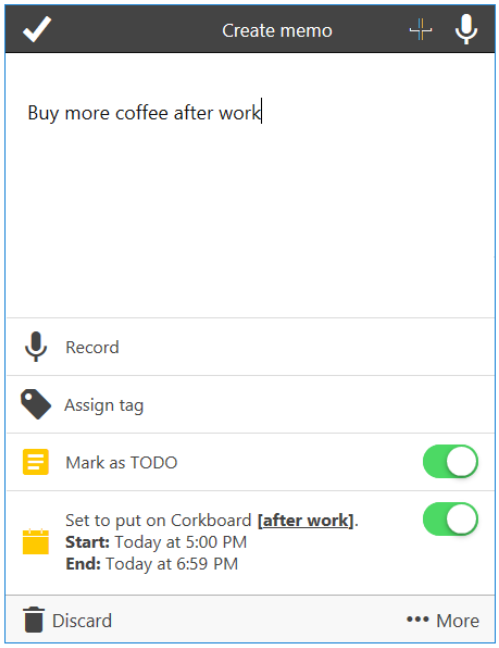

Figure 11. Participants can include temporal expressions in the text of their memos, or can click on the scheduling panel (bottom) to pick times or dates from a list of suggestions or from a calendar.

Procedure

Participants were recruited by email, within the same large IT organization as the earlier survey. Participants were screened for past experience with Android smart phones, ensuring that they were familiar with the UI conventions of the device used in the study. In total, 14 participants were interviewed (5 female, mean age = 31 years), and each received a $5 food voucher as remuneration.

Upon arriving to the interview, participants were presented a brief video tutorial demonstrating the features of the probe system. Participants were then tasked with inputting four of their own personal to-do items they wished to remember. After participants entered the four notes, the interviewer (an author on this paper) revisited each item in turn, and asked participants to comment on the appropriateness of the system's interpretations. The interviewer also asked participants if they would like to receive additional notifications or reminders for each item. These discussions were recorded and transcribed. Iterative open-coding techniques were applied to identify the common themes.

At the end of the interview, participants completed a questionnaire which included both open-ended questions about the usability of the system and a balanced subset[2] of Likert items from the System Usability Scale (SUS) [5]. Individual Likert items were scored on a 5-point scale (strongly disagree = 1, strongly agree = 5). Given our research focus, and the limited implementation of the design probe, this exit questionnaire served as a check to ensure that the system was of sufficient quality and completeness to effectively ground the discussions – we did not want the interviewees dwelling on software bugs or missing features.

[2]We used four positively-worded and four negatively-worded items.

Results and Discussion

General impressions

All 14 participants successfully accomplished their input tasks. Of the 56 items generated in total, 23 items contained an in-line temporal expression, 9 items were assigned times from the list of suggestions, and 6 items were assigned times via the calendar or time-picker widgets. The remaining 18 items lacked any explicit temporal cues, and defaulted to "today" (i.e., the item was scheduled to remain displayed for the rest of the day). In-line temporal expressions ranged from simple (e.g., "*Freshman orientation September 25th*", P4), to complex (e.g., "*Set up Tokyo meeting with (a business) early next week for an evening meeting*", P8).

The virtual assistant also received high scores on specific SUS usability items: 12 participants agreed that the system was easy to use (M=3.4, SD=0.82), 10 participants agreed that they felt confident using the system (M=3.2, SD=0.77), and all 14 participants agreed that most people would learn the system quickly (M=3.4, SD=0.82). Conversely, 9 participants *disagreed* with the statement asserting that the system was cumbersome (M=1.9, SD=0.88), and all 14 participants disagreed with the statement asserting that they would require technical support to use the system (M=1.3, SD=0.47).

Finally, we report participants' favorite and least favorite features. 7 of the 14 participants listed the ability to dictate memos and to-dos as their favorite feature (e.g., "*simple and easy voice input*"), while 6 participants listed the system's ability to detect in-line temporal expressions as their favorite feature (e.g., "*it was smart about figuring out the time starts and ends*"). Conversely, the list of least favorite features was varied, and many concerns pertained to specifics of the UI (e.g., "*I was expecting a save button rather than a back button*")

Themes

Common themes that emerged from the participants' comments on the system are summarized and discussed below.

Implied flexibility. A prevalent theme that emerged from the interviews was that many items carried an implied flexibility in scheduling. For example, after inputting the item "*Buy Christmas presents*", participant 1 (P1) noted that the system should display the item "*somewhere in December*," and that "*(he) doesn't care if that's the first, second, third, tenth, twentieth (of December), but it should be before the 24th*." Likewise, after inputting the note "*revisit will in 5 years*," P14 was surprised by the system's literal interpretation of August 26, 2021 – exactly five years from the day the entry was created. The participant noted that the item did not require a specific date, and that the appropriate granularity was at the level of months. These sentiments were best summarized by P1, who noted:

> "*I want the reminder to be as imprecise as my time instruction. If I say 11am and 3 minutes and 10 seconds, I expect a very precise reminder that notifies me maybe a minute before that, but if I say 'next month', then I expect it to remind me a few days ahead, even a few weeks ahead of it.*"

Implied constraints. While the former items carried implicit flexibility, others required that the system consider unstated constraints. For example, when P4 input the to-do item "*Remember to pack presents tomorrow*," the assistant responded by offering to display the item on the Corkboard from 12:00am to 11:59pm on the day following the interview. When asked if this response was appropriate, P4 responded "*So probably what I should have put in here was 'tomorrow night' since that's exactly what I meant, or what would have been more useful to me.*" Likewise, after inputting the to-do "*do corporate training tomorrow*", P2 noted that "*business hours would have made more sense.*" In these particular cases, participants expected the system to understand the natural constraints of the tasks: packing is done at home and on-the-job training is done at work.

In several instances, implied constraints co-occurred with implied flexibility. For example, after entering "*Go to grocery store on Sunday*," P5 responded:

> *"Ideally it should look at my calendar and figure out when is a good time to go. Or, actually, several different time slots. It's kind of a vague thing, right."*

In this latter situation, the participant wanted the system to recognize both that the task was flexible, and that it required about two hours of free time to complete.

Implied task preparation time. In other cases, system interpretations were problematic because they failed to capture preparation times. For example, upon inputting the phrase "*Make pumpkin pie for thanksgiving*," P5 noted that "*I would need ingredients so I would probably want (a reminder) a few days before.*" Likewise, when P6 input the item "*Pack bags for Sunday,*" she noted that her flight departed Sunday morning, and that her bags would have to be packed well in advance.

These two examples illustrate a common trend: prefix words like "by" or "for" often indicate a planning activity, and thus could prove to be useful signals that the system should consider preparation time.

However, there were also cases where the need for preparation was implicit. Upon inputting the note "*Remember Mom's birthday June 25,*" P3 noted he wanted the information to appear at least a week before so that he would have time to find a present. Likewise, P4 noted that, for any item occurring in the early morning, she would like a reminder "*sometime before bed, saying 'hey you've got things early tomorrow'*." This evening notification would serve to "*remind (her) to set the alarm a little earlier*."

Complex expressions. Several situations arose in which the language was complex or ambiguous. For example, P13 recorded the memo "*Follow up with (a friend) about dinner plans either Saturday or Sunday.*" Upon seeing the item was scheduled for display on the following Saturday, P13 remarked that the interpretation was incorrect, then explained: "*It's not like I was telling it to mark it for Saturday or Sunday, I was telling it I need to follow up with her about Saturday or Sunday.*" Likewise, upon inputting "*Buy tickets for the concert next month*," P9 was disappointed the system misinterpreted "next month" as the time when tickets should be purchased, rather than the time at which the concert was occurring.

In summary, we found that participants often expected our virtual assistant to recognize implicit flexibilities, constraints, and preparation activities. Consequently, even when items contained precise temporal expressions (e.g., specific dates), they often exhibited characteristics similar to ITEs. As such, it is especially important that virtual assistants be designed to handle temporal uncertainty. To this end, we present design implications in the next section.

DESIGN IMPLICATIONS

Throughout the discussions in the previous sections, there are several recurring themes that are indicative of how virtual assistants should behave when addressing ITEs. These expected behaviors are summarized below.

Respect uncertainty. Both our survey and interview studies indicate that ITEs can serve various communicative purposes, such as to convey task unimportance and schedule flexibility (Figure 4). Therefore, virtual assistants should avoid early resolution of uncertainty, unless absolutely necessary. Consider, for example, a scenario in which a person marks a task or memo for "*later this week*." At the moment of this interaction the person may lack the information (or the time) required to more precisely schedule the item. If, in this instant, the assistant prompts the user to input a precise time, then it is likely to appear tone-deaf to this reality.

Recognize uncertainty. Our analysis of the Avocado email corpus reveals that ITEs are frequent and diverse – 36.2% of observed time expressions were imprecise, and nearly 80% were observed 3 or fewer times (Figure 5 and Table 2). Complex, "long-tail" expressions were also observed to occur in the interview study. This is notable because, in contrast to the

Avocado email corpus, interviewees formed their utterances with the expectation that they would be interpreted by a virtual assistant. If virtual assistants are to be designed to respect ITEs, they must be able to generalize such that they recognize expressions rarely seen in pre-existing corpora.

Embrace flexibility. Designing virtual assistants to delay the resolution of uncertainty, as above, raises an obvious question: When should temporal uncertainty be resolved? A reasonable strategy is to have virtual assistants embrace any inherent flexibility, and they should do so opportunistically. Continuing the previous "later this week" example, unless "this week" is drawing to an end, there is likely plenty of time for the virtual assistant to seek additional clarification as part of some later maintenance task. For instance, the assistant might ask for scheduling guidance the next time the user opens their calendar, or as part of daily briefings delivered at convenient down-times (e.g., at the start or end of the day). These low-urgency situations may even lend themselves well to crowd-sourcing scenarios, where crowd-workers can leverage common knowledge to identify implicit details and constraints.

Notify intelligently. In situations where uncertainty is high and urgency is low, virtual assistants should strive to deliver information using non-intrusive methods. For example, the lock screen of a phone or the watch face of a wearable device could provide convenient ambient displays for this type of information – especially given their associations with time-keeping. Such a mechanism can be supported by proactive behavior models proposed in existing literature [4, 38].

Leverage implicit knowledge. Our interview study indicates that users expect virtual assistants to possess more real-world knowledge than they currently do. Overall, we advocate making advances in the following three areas: (1) recognizing cultural or personal milestones, such as holidays and birthdays; (2) identifying natural constraints (e.g., packing a suitcase is likely done at home; baking a pie likely takes a few hours, etc.); and, (3) learning individual users' behavior patterns, such as commute schedule or sleep patterns.

DISCUSSION AND LIMITATIONS

In this paper, we explored existing practices, expectations, and preferences surrounding the use of temporal expressions in interactions with virtual assistants. Our analyses employed multiple complementary methodologies, and considered data contributed by 631 individuals.[3] Whenever possible, our reporting highlighted themes common to two or more methods or data sets to detail the opinions of the specified demographic.

That being said, we caution readers against overly generalizing these findings. Our survey respondents and interview participants, together with the individuals represented by the Avocado email corpus, were employees of U.S.-based technology companies at the time of data collection. While our data sets cover a variety of job roles, it remains to be shown how well these findings apply to a more general U.S. population, and, especially, to a more general global population. We are particularly interested in learning how cultural milestones (e.g., holidays), attitudes (e.g., on punctuality) and norms (e.g., the time meals are consumed) may impact the issues considered in this paper. We leave these investigations to future research.

It is also worth noting that the situations presented in the survey and interviews focus on task management scenarios (i.e., to-dos and memos). These scenarios are common: Graus et al. reported that 576,080 time-based reminders were created between January and February 2015, in a subsample of 92,264 US-resident users of Cortana-enabled mobile devices [12]. However, these calendaring/reminder scenarios are by no means the only tasks performed with virtual assistants. In 2014, Jiang et al. reported that such interactions accounted for 18.2% of Cortana's primary operations (i.e., all operations excluding chit-chats and generic Web searches) [13]. In the future, it will be important to consider how temporal expressions are manifest in other scenarios involving virtual assistants.

Finally, it is worth considering how legacy bias [21] – one's past experience with virtual assistants – may influence participant responses. The specific concern is that virtual assistants differ in syntax and features. Our demographic collection was not sufficiently fine-grained to allow us to compare the user populations of specific virtual assistants. However, our survey did ask if participants had *any* past experience with virtual assistants. Our subsequent analysis failed to find any significant differences between respondents who reported previous experience with virtual assistants, and those who did not (Mann–Whitney U, $\alpha = 0.05$, with, and without, Bonferroni correction). Differences between the user populations of competing virtual assistants would be interesting, if found; but, we leave this to future work.

CONCLUSION

We have investigated the motivation, manifestation, and expectation surrounding the use of imprecise temporal expressions in people's communication with virtual assistants. We have found that the primary motivations of using imprecise time expressions involve low level of commitment, low task priority, or external dependencies of the tasks in question. We also use text corpus analysis to find that imprecise temporal expressions have high variability and many exhibit temporal dependencies. Additionally, our interviews reveal that people expect the virtual assistants to be able to handle the implied flexibility, implied constraints, and task preparation activities from people's imprecise temporal expressions. Finally, we draw design implications for future virtual assistants; they should: (1) respect uncertainty by delaying early resolution unless absolutely necessary; (2) recognize uncertainty by supporting the long-tail of imprecise temporal expressions; (3) embrace flexibility by considering engaging with the user during down-times or alternative metaphors (such as our Cork-board) that enable uncertainty to be resolved at a later time; (4) adopt alternative strategies to deliver reminder notifications appropriate to the uncertainty of a time expression; and (5) leverage implicit personal information or world knowledge through sources such as calendars, external event knowledge (e.g., traffic), and predicted or user-entered preferences.

[3] Includes 338 survey respondents, 279 email account holders in the Avocado data set, and 14 interview participants.

REFERENCES

1. Gabor Angeli, Christopher D. Manning, and Daniel Jurafsky. 2012. Parsing Time: Learning to Interpret Time Expressions. In *Proc. NAACL HLT '12*. ACL, 446–455.
2. Jacob Bank, Zachary Cain, Yoav Shoham, Caroline Suen, and Dan Ariely. 2012. Turning personal calendars into scheduling assistants. In *Proc. CHI '12*. ACM, 2667–2672.
3. Frank Bentley, David A Shamma Joseph 'Jofish' Kaye, and John Alexis Guerra-Gomez. 2016. The 32 Days Of Christmas: Understanding Temporal Intent in Image Search Queries. In *Proc. CHI '16*. ACM, 5710–5714.
4. Pauline M. Berry, Melinda Gervasio, Bart Peintner, and Neil Yorke-Smith. 2011. PTIME: Personalized Assistance for Calendaring. *ACM TIST* (2011), 40:1–40:22.
5. John Brooke and others. 1996. SUS-A quick and dirty usability scale. *Usability Evaluation in Industry* 189, 194 (1996), 4–7.
6. Angel X Chang and Christopher D Manning. 2012. SUTime: A library for recognizing and normalizing time expressions. In *Proc. LREC '12*. 3735–3740.
7. Richard W DeVaul, Brian Clarkson, and others. 2000. The memory glasses: towards a wearable, context aware, situation-appropriate reminder system. In *CHI 2000 Workshop on Situated Interaction in Ubiquitous Computing*. ACM.
8. Anind K Dey and Gregory D Abowd. 2000. CybreMinder: A context-aware system for supporting reminders. In *International Symposium on Handheld and Ubiquitous Computing*. Springer, 172–186.
9. Didier Dubois and Henri Prade. 1989. Processing fuzzy temporal knowledge. *IEEE Transactions on Systems, Man, and Cybernetics* 19, 4 (1989), 729–744.
10. Lisa Ferro, Laurie Gerber, Inderjeet Mani, Beth Sundheim, and George Wilson. 2005. TIDES 2005 standard for the annotation of temporal expressions. (2005).
11. Michael Freed, Jaime G Carbonell, Geoffrey J Gordon, Jordan Hayes, Brad A Myers, Daniel P Siewiorek, Stephen F Smith, Aaron Steinfeld, and Anthony Tomasic. 2008. RADAR: A Personal Assistant that Learns to Reduce Email Overload. In *Proc. AAAI '08*. 1287–1293.
12. David Graus, Paul N. Bennett, Ryen W. White, and Eric Horvitz. 2016. Analyzing and Predicting Task Reminders. (2016), 7–15.
13. Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. 2015. Automatic Online Evaluation of Intelligent Assistants. In *Proc. WWW '15*. ACM, 506–516.
14. Ece Kamar and Eric Horvitz. 2011. Jogger: models for context-sensitive reminding. In *Proc. AAMAS '11*. IFAAMAS, 1089–1090.
15. Mik Lamming and Mike Flynn. 1994. Forget-me-not: Intimate computing in support of human memory. In *Proc. FRIEND21, '94 Int. Symp. on Next Generation Human Interface*. 4.
16. Kenton Lee, Yoav Artzi, Jesse Dodge, and Luke Zettlemoyer. 2014. Context-dependent Semantic Parsing for Time Expressions. In *ACL*. 1437–1447.
17. Pamela J Ludford, Dan Frankowski, Ken Reily, Kurt Wilms, and Loren Terveen. 2006. Because I carry my cell phone anyway: functional location-based reminder applications. In *Proc. CHI '06*. ACM, 889–898.
18. Ewa Luger and Abigail Sellen. 2016. Like Having a Really Bad PA: The Gulf between User Expectation and Experience of Conversational Agents. In *Proc. CHI '16*. ACM, 5286–5297.
19. Inderjeet Mani and George Wilson. 2000. Robust temporal processing of news. In *Proc. ACL 2000*. ACL, 69–76.
20. Reed Martin and Henry Holtzman. 2011. Kairoscope: managing time perception and scheduling through social event coordination. In *Proc. CHI '11*. ACM, 1969–1978.
21. Meredith Ringel Morris, Andreea Danielescu, Steven Drucker, Danyel Fisher, Bongshin Lee, m. c. schraefel, and Jacob O. Wobbrock. 2014. Reducing Legacy Bias in Gesture Elicitation Studies. *Interactions* 21, 3 (2014), 40–45.
22. Karen Myers, Pauline Berry, Jim Blythe, Ken Conley, Melinda Gervasio, Deborah L McGuinness, David Morley, Avi Pfeffer, Martha Pollack, and Milind Tambe. 2007. An intelligent personal assistant for task and time management. *AI Magazine* 28, 2 (2007), 47.
23. Douglas Oard, William Webber, David Kirsch, and Sergey Golitsynskiy. 2015. Avocado Research Email Collection. LDC2015T03. DVD. Philadelphia: Linguistic Data Consortium. (2015).
24. Tim Paek and Eric Horvitz. 1999. Uncertainty, utility, and misunderstanding: A decision-theoretic perspective on grounding in conversational systems. In *AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems, Cape Cod, MA*.
25. Joelle Pineau, Michael Montemerlo, Martha Pollack, Nicholas Roy, and Sebastian Thrun. 2003. Towards robotic assistants in nursing homes: Challenges and results. *Robotics and autonomous systems* 42, 3 (2003), 271–281.
26. James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and others. 2003. The timebank corpus. In *Corpus linguistics*, Vol. 2003. 40.
27. Anand S Rao and Michael P Georgeff. 1992. An abstract architecture for rational agents. *KR* 92 (1992), 439–449.

28. Ioannis Refanidis and Neil Yorke-Smith. 2010. A constraint-based approach to scheduling an individual's activities. *ACM Transactions on Intelligent Systems and Technology (TIST)* 1, 2 (2010), 12.
29. Frank Schilder and Christopher Habel. 2003. Temporal Information Extraction for Temporal Question Answering. In *New Directions in Question Answering*. 35–44.
30. Steven Schockaert and Martine De Cock. 2008. Temporal reasoning about fuzzy intervals. *Artificial Intelligence* 172, 8 (2008), 1158–1193.
31. Steven Schockaert, Martine De Cock, and Etienne E Kerre. 2007. Qualitative Temporal Reasoning about Vague Events. In *IJCAI*. 569–574.
32. Timothy Sohn, Kevin A Li, Gunny Lee, Ian Smith, James Scott, and William G Griswold. 2005. Place-its: A study of location-based reminders on mobile phones. In *Proc. UbiComp '05*. ACM, 232–250.
33. Jannik Strötgen and Michael Gertz. 2010. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In *Proc. International Workshop on Semantic Evaluation*. ACL, 321–324.
34. Éric Taillard, Philippe Badeau, Michel Gendreau, François Guertin, and Jean-Yves Potvin. 1997. A tabu search heuristic for the vehicle routing problem with soft time windows. *Transportation science* 31, 2 (1997), 170–186.
35. John Thangarajah, James Harland, David Morley, and Neil Yorke-Smith. 2008. Suspending and resuming tasks in BDI agents. In *Proc. AAMAS '08*. IFAAMAS, 405–412.
36. Hegler Correa Tissot. 2016. *Normalisation of imprecise temporal expressions extracted from text*. Ph.D. Dissertation. University of Paraná.
37. Giorgio Vassallo, Giovanni Pilato, Agnese Augello, and Salvatore Gaglio. 2010. *Phase Coherence in Conceptual Spaces for Conversational Agents*. New York, NY, USA: Wiley, IEEE Press.
38. Neil Yorke-Smith, Shahin Saadati, Karen L Myers, and David N Morley. 2012. The design of a proactive personal agent for task management. *International Journal on Artificial Intelligence Tools* 21, 01 (2012), 1250004.
39. Luis Zabala, Cristina Perfecto, J Unzilla, and A Ferro. 2001. Integrating automatic task scheduling and Web-based agenda in a virtual campus environment. In *Proc. ITHET '01*.
40. Lotfi A Zadeh. 1994. Soft computing and fuzzy logic. *IEEE software* 11, 6 (1994), 48.
41. H-J Zimmermann. 2010. Fuzzy set theory. *Wiley Interdisciplinary Reviews: Computational Statistics* 2, 3 (2010), 317–332.