Question Answering over Linked Data with Vague Temporal Adverbials

David Maria Schmidt[1,*][a], Svenja Kenneweg[1,*][b], Julian Eggert[2][c], Jörg Deigmöller[2][d] and Philipp Cimiano[1][e]

[1]*Semantic Computing Group, CITEC, Technical Faculty, Bielefeld University, Bielefeld, Germany*

[2]*Honda Research Institute Europe, Offenbach, Germany*

Keywords: Question Answering over Linked Data, Vague Temporal Adverbials, Compositional Question Answering.

Abstract: Vague temporal adverbials are common in human communication but most question answering over linked data (QALD) approaches only work with exact time points. We present a QALD system that interprets vague temporal adverbials (e.g., "just", "recently") using a factorized probabilistic model. Building on NeoDUDES, an existing QALD approach, we map vague temporal adverbials to time intervals via empirically grounded Gaussian functions and generate SPARQL queries with temporal filters, enabling compositional interpretation of questions involving vagueness. Evaluated on a knowledge graph based on real-world smart home data, our system shows strong performance.

1 INTRODUCTION

Humans often rely on vague temporal adverbials such as *just*, to describe past events when their exact happening time is irrelevant (Van Jaarsveld and Schreuder, 1985; May et al., 2021). Unlike explicit references (e.g., *on 17.01.2024 at 13:00*), these adverbials lack a precise point in time but still convey a shared intuitive meaning (e.g., "I just took a bath" usually implies earlier today, while "I just cleaned the house" could mean yesterday or even two days ago.).

Question Answering over Linked Data (QALD) maps natural language questions into SPARQL queries that can be executed over a knowledge graph to compute corresponding answers. Despite the ubiquity of vague expressions in natural language, most QALD systems for temporal knowledge graphs solely focus on queries anchored to exact temporal points or intervals (Jia et al., 2021; Kannen et al., 2023; Chen et al., 2022; Sharma et al., 2023; Huang et al., 2024). This mismatch between human use of vague temporal adverbials and the capabilities for their interpretation by current question answering systems limits the practical utility of current QALD systems.

To address this gap, we extend the NeoDUDES QALD system (Schmidt et al., 2025), adding support for vague temporal adverbials. We added a configurable reference time as well as a lexicon comprising vague temporal adverbials and events. Additionally, we modified the DUDES creation and SPARQL generation modules in order to translate vague temporal adverbials and corresponding events into query clauses guided by the factorized probabilistic model proposed by Kenneweg et al. (Kenneweg et al., 2025a), which we call `FuzzyLLI` ("Fuzzy probabiListic adverbiaL Interpretation"). Through a special predicate `vaguetemp`, the NeoDUDES pipeline calls `FuzzyLLI`, provides the event and respective adverbial, and gets back a crisp interval as an interpretation of the adverbial. Adding the reference time to this interval yields the bounds for the corresponding `FILTER` statements. This crisp interval, determined by `FuzzyLLI` (Kenneweg et al., 2025a), represents the most likely time span - in minutes relative to the reference time - during which the event described by the vague temporal adverbial is assumed to have occurred with a probability exceeding a predefined threshold. As the original model only accounts for six different events, we generalize the model using a decision-tree classifier based on the typical duration and frequency of the corresponding event. This allows our model to generalize to any daily event with a duration typically expressed in minutes or hours, with the limitation that

[a] https://orcid.org/0000-0001-7728-2884
[b] https://orcid.org/0009-0002-3025-7563
[c] https://orcid.org/0000-0003-4437-6133
[d] https://orcid.org/0009-0007-5931-6973
[e] https://orcid.org/0000-0002-4771-441X
* Equal contribution

Schmidt, D. M., Kenneweg, S., Eggert, J., Deigmöller, J. and Cimiano, P.
Question Answering over Linked Data with Vague Temporal Adverbials.
DOI: 10.5220/0013778400004000

In *Proceedings of the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2025) - Volume 2: KEOD and KMIS*, pages 164-174
ISBN: 978-989-758-769-6; ISSN: 2184-3228

the event's duration needs to be manually added to `FuzzyLLI`.

We evaluate our system in a controlled setting using the CASAS smart home datasets (Cook et al., 2013), specifically the annotated twor.2010 dataset (Cook and Schmitter-Edgecombe, 2009), which includes data collected from two residents performing thirteen common household events. After preprocessing, we refine the dataset to include eleven distinct events and transform it into a temporal RDF knowledge graph (KG). Based on this KG, we generate our evaluation dataset, containing 2,780 natural language questions paired with ground truth answers. The questions are created using templates spanning four categories, organized as follows:

1. **Did** - e.g., *Did Tom sleep some time ago?*
2. **What** - e.g., *What did Mary do a long time ago?*
3. **What happened** - e.g., *What happened recently?*
4. **Who** - e.g., *Who has just watched TV?*

In short, our contributions are as follows:

- We introduce a novel extension to an existing QALD system that enables the interpretation of questions containing vague temporal adverbials.
- We propose a complete pipeline integrating Lemon (McCrae et al., 2011) lexica, configurable reference times, and enriched SPARQL generation to support vague temporal QALD.
- We extend a model for vague temporal adverbials to handle more events, using a decision-tree based on event frequency and duration.
- We evaluate our system on a newly constructed dataset comprising 2,780 questions and answers based on real-world household events.

The exact match rate of our system, i.e., the number of exact matches divided by the total number of benchmark items, is very promising, achieving scores between 0.85 and 0.91 for the best-performing query selection model. However, these results are based on a number of assumptions that limit the generalizability of the approach to other domains and events, which we discuss in detail in the paper. In particular, only events which happen daily and have a duration typically measured in minutes or hours are supported by the system, with the duration requiring manual classification in advance. Further, the approach relies on lexical entries for all relevant words. Future work concerns how well our system performs on other types of questions. The source code and data used in our experiments are available at Zenodo: https://doi.org/10.5281/zenodo.16893293.

2 RELATED WORK

2.1 (Temporal) Tagger

Temporal taggers identify and normalize temporal expressions in text by mapping them to standardized formats such as ISO. A key markup language for annotating temporal information is TimeML (Pustejovsky, 2005), with ISO-TimeML as its revised and interoperable revision (Pustejovsky et al., 2010). A simpler alternative is TIE-ML (Cavar et al., 2021).

When looking at available temporal taggers, there exist both rule-based and language model-based approaches. For example, HeidelTime (Strötgen and Gertz, 2010) and SUTIME (Chang and Manning, 2012) are two well-known rule-based systems, whereas (Lange et al., 2023) and (Schilder and Habel, 2001) are based on masked language modeling. Additionally, there also exist some taggers supporting vague expressions (May et al., 2021). In contrast, the focus of this work is to evaluate and interpret vague temporal adverbials w.r.t. a knowledge graph to obtain relevant information using SPARQL queries.

2.2 (Temporal) Ontologies and Reasoners

There are a multitude of ontologies supporting the modeling of temporal and also vague concepts. There exist in particular different ontologies with a focus on temporal aspects in the context of the Web Ontology Language (OWL). Two examples are OWL-Time[1] and FuzzyOWL (Stoilos et al., 2005). OWL-Time is an OWL-2 DL ontology supporting the modeling of temporal concepts that can be used to describe temporal properties of resources. We use OWL-Time in our knowledge graph to specify the interval in which an event happened. FuzzyOWL handles uncertainty and vagueness by introducing a degree value that describes to which degree a certain concept applies to a resource, like "tall" to a person. This can, however, not be applied to vague temporal adverbials easily, as their degree value would not just depend on the adverbial and the event, but also on the temporal distance from a non-static reference time point. Since generating values for all possible reference times is not feasible, we follow a more dynamic approach in this paper, using `FuzzyLLI`. Moreover, the DUL (DOLCE+DnS Ultralite) ontology also covers certain temporal aspects. We use `dul:hasAgent` in our knowledge graph to model who performed an event. In addition, there is also a wide variety of OWL reasoners, such as ELK

[1] https://www.w3.org/TR/owl2-overview/

(Kazakov et al., 2014), RDFox (Nenov et al., 2015), and ldfu (Käfer and Harth, 2018). Few reasoners, however, are capable of processing vague predicates, e.g., DeLorean (Bobillo et al., 2012) and fuzzyDL (Bobillo and Straccia, 2016).

2.3 Vague Expressions

Vagueness arises when an expression has borderline cases – instances where it neither clearly applies nor clearly fails to apply. For example, the adjective *tall* is vague because the boundaries for when a person counts as *tall* are not fixed. Vague expressions are typically context-dependent: *Young* may describe a baby of a few months or an adult at the age of twenty (Damerau, 1977). While most of the literature focuses on vagueness in adjectives (Kamp and Sassoon, 2016; Solt and Gotzner, 2012), less attention has been given to vague temporal adverbials, such as *recently* or *a long time ago*. These adverbials refer to past events relative to the time of utterance, but leave the exact time this event took place underspecified.

An exception is the work by (Kenneweg et al., 2024), who performed an online survey to measure how native English speakers interpret adverbials such as *recently*, or *long time ago* in relation to different types of events that have occurred a certain time ago. For example, participants rated the appropriateness of statements like "Tom had his birthday recently" when the birthday occurred one day ago. The results allow to quantify how likely one of their empirically-measured adverbials is to be used to describe one of their empirically-measured events that happened t units of time ago.

(Kenneweg et al., 2025a) have also proposed a model that captures these interpretations. Crucially, the authors demonstrated that Large Language Models perform poorly at this task: They struggle to identify the correct time ranges of events described by vague temporal adverbials, when compared with the humans' interpretations from the above empirical work (Kenneweg et al., 2024). This highlights the need for an explicit model of the meaning of temporal adverbials. The model is described in detail in Section 3.2, as well as the extension supporting normalization across a broader range of events in Section 4.3.

2.4 Temporal Knowledge Graph Question Answering

There already exist various QALD approaches that can deal with temporal knowledge graphs (see, e.g., (Su et al., 2024) for an overview). However, these approaches usually focus on temporal predicates with exact boundaries or precise relationships between intervals, either stated explicitly (Jia et al., 2021) or implicitly (Kannen et al., 2023; Chen et al., 2022; Sharma et al., 2023; Huang et al., 2024). Temporal relationships typically captured include "before"/"after" and "during"/"include" (Neelam et al., 2021; Mavromatis et al., 2022; Jiao et al., 2023; Chen et al., 2022). However, to the best of our knowledge, none of these approaches deals with vague temporal expressions, i.e., temporal predicates that do not have precisely-defined boundaries.

3 METHOD

3.1 Question Answering System

We present a QALD system supporting different vague temporal adverbials for different kinds of events. To do so, we build on the work of (Schmidt et al., 2025), extending the NeoDUDES QALD system. Due to its modular, compositional and lexical knowledge-based nature, it is very well-suited for adding support for vague temporal expressions.

The NeoDUDES pipeline works by first applying dependency parsers to the input question. The resulting dependency trees are then compacted by merging different tree nodes based on a set of heuristics, e.g., merging determiners into their parent nodes. This is done to facilitate ontology matching in the next step, during which candidate entities and properties from the target ontology are assigned to each tree node. This is achieved by utilizing different data sources, most importantly lexical entries that bridge the lexical gap between natural language and ontology resources.

As ambiguities may arise throughout all of these steps, the approach accounts for all possible combinations in those cases. As this results in a large number of possible interpretations, the tree scoring step ranks the available ontology-matched trees by how promising they appear to be. More precisely, the scores include how many nodes there are compared to the unmerged dependency tree, how many nodes have successfully been matched with an ontology resource, and how well those resources match the node. The resulting order determines which candidate is further investigated first, thus focusing on the most promising paths first, without discarding other possibilities.

Afterwards, based on the assigned ontology resources, *Dependency-based Underspecified Discourse Representation Structures (DUDES, (Cimiano, 2009; Cimiano et al., 2014))* are created. A DUDES consists of three main parts: i) a list of logical expressions representing the relations between

the matched ontology resources, ii) a main variable used during composition of two DUDES in combination with iii) selection pairs, indicating variables in a DUDES that are not bound to a fixed value or already unified with a variable from another DUDES.

A formal composition operator is defined for DUDES, allowing to compose two DUDES into a single resulting DUDES that represents the combined meaning of both input DUDES. Ultimately, this allows the composition of all DUDES of a tree into a unified representation of the input question's meaning. This is done in the DUDES composition step.

Based on the composed final DUDES of a candidate tree, a SPARQL query can be generated straightforwardly. The basic triple patterns are generated by using the Z3 SMT solver (de Moura and Bjørner, 2008) to determine which variables are bound to a fixed value and which are free and have to be translated to SPARQL variables accordingly. More sophisticated SPARQL syntax, e.g., `FILTER`, is handled by introducing special properties that are processed separately and trigger the introduction of the corresponding SPARQL syntax. As ambiguities can also arise in the steps after the tree scoring, there may be multiple candidate SPARQL queries even for a single candidate tree. Therefore, the SPARQL selection step, similarly to the tree scoring step, selects the final SPARQL query returned as final output using an LLM-based approach. The final SPARQL query can then be evaluated against the target knowledge graph to retrieve the actual answer.

3.2 Modeling Vague Temporal Adverbials

In order to enable the NeoDUDES QALD system (see Section 3.1) to handle questions involving vague temporal adverbials, such as *"Did Tom just brush his teeth?"*, it is key to determine the most likely temporal interval, relative to a reference time, during which the event "brushing teeth" - described by the adverbial "just" - took place. As discussed in Section 2.3, the interpretation of temporal adverbials depends on their comparison class. Consequently, we adopt a factorized modeling approach, inspired by Frege's principle of compositionality (Frege, 1953), which posits that *"the meaning of a complex predicate can be modeled via the meaning of its parts and how they are composed together"*.

In our case, relevant components are the vague temporal adverbial (e.g., "just," "recently") and its comparison class - the event (e.g., "brushing teeth", "sleep"). We base our approach on the factorized compositional model proposed by (Kenneweg et al., 2025a), `FuzzyLLI`, which estimates the probability that a human would use temporal adverbial *Adv* to describe an event *Ev* that occurred t time units in the past: $P_{Adv}(P_{Ev}(t))$. They model the probability $P_{Ev}(t)$ by using the cumulative distribution function of a Gaussian distribution:

$$P_{Ev}(t) = \frac{1}{2}\left(erf\left(\frac{t}{\sqrt{2}\sigma_e}\right)+1\right) \tag{1}$$

In this equation, the parameter σ_e depends on the event. A smaller σ_e means the probability increases faster together with the temporal distance t to the reference time. A larger σ_e, in turn, leads to a slower probability increase. The probability associated with a temporal adverbial P_{Adv} is modeled with a normalized Gaussian distribution:

$$P_{Adv} = \exp\left(-\frac{1}{2}\left(\frac{x-\mu_a}{\sigma_a}\right)^2\right) \tag{2}$$

In this function, μ_a and σ_a are the mean and standard deviation, respectively, which are unique for each temporal adverbial. Consider again the question *"Did Tom just brush his teeth?"* from the beginning. In this example, the adverbial is *just* and the event *brushing teeth*. To identify the most plausible temporal interval during which *brushing teeth* could *just* have taken place, we calculate $P_{Adv}(P_{Ev}(t))$, which results in different possibilities depending on t. A chosen threshold θ defines the lower and upper bounds of the temporal interval during which *brushing teeth* could plausibly *just* have taken place. More precisely, the bounds are the minimal and maximal t for which $P_{Adv}(P_{Ev}(t))$ is above or equal to θ.

To determine σ_e, μ_a and σ_a for specific events and adverbials, the parameters of `FuzzyLLI` have been empirically fitted by (Kenneweg et al., 2025a) using data from a previous study (Kenneweg et al., 2025b), as described in Section 2.3. The dataset (Kenneweg et al., 2025b) includes four temporal adverbials (*just*, *recently*, *some time ago*, *long time ago*) and six events (*brushing teeth*, *birthday*, *vacation*, *marriage*, *sabbatical*, and *year abroad*), thus limiting the existing model to these specific events and adverbials. Given that the current model developed by (Kenneweg et al., 2025a) does not directly support the events present in our dataset, we extended their approach accordingly, as described in Section 4.3.

3.3 Extensions to NeoDUDES

In principle, there are three main ways of handling vague temporals in a QALD setting. First, one can include the temporal information in the SPARQL query by transforming the adverbial into a time interval depending on the event in question. This interval can

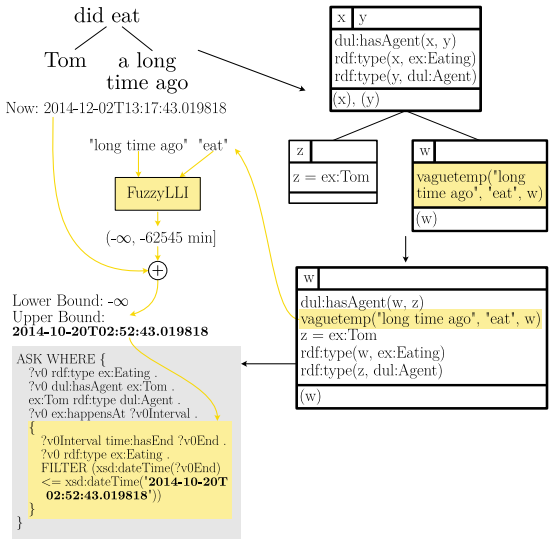

Figure 1: Illustration of initial tree representation over DUDES creation and composition to final SPARQL query for "Did Tom eat a long time ago?". Yellow parts represent new behavior for vague temporal expressions.

then be included into the query within a `FILTER` statement, thus turning the vague temporal adverbial into a crisp interval. Another way is to generate a general SPARQL query, returning all potential candidates without the vague temporal constraint and handling the interpretation of vagueness in a separate step, post-processing the query bindings. This has the advantage of remaining flexible w.r.t. a concrete interpretation of the respective vague temporal adverbial. However, this extra step substantially differs from the typical QALD setting, where all information is usually condensed in a SPARQL query. Finally, a third way to address vagueness would be by adding built-in functions to the SPARQL query engine. These built-in functions would be evaluated as part of the reasoning procedures. Although this allows both to remain flexible w.r.t. interpretation and have all information as part of the SPARQL query, it also needs a specialized SPARQL endpoint supporting the definition and evaluation of such built-in functions.

In this paper, we decided to follow the first approach and extend the existing NeoDUDES pipeline in order to support the interpretation of vague temporal expressions. First, we pass a reference time point to the pipeline, such that reasoning w.r.t. a specific (and adaptable) definition of "now" is possible. Additionally, we provide a small set of 17 lexical entries for the target knowledge graph described in Section 4.1.2. This lexicon mainly consists of two categories of entries. First, the supported events (e.g., "eat" or "sleep") are bound to the `dul:hasAgent` property. Second, there exist entries for the supported temporal adverbials (e.g., "recently" or "a long time ago"), identifying them as temporal adverbials. The entries allow the pipeline to recognize terms relevant for the `FuzzyLLI` module and to forward them accordingly to build `FILTER` statements for the target KG.

The extension of the pipeline mainly involved the modification of the DUDES generation process, adding a semantic representation involving a special property `vaguetemp` for vague temporal adverbials. The property is added when vague temporal adverbials are encountered in the dependency tree, i.e., when a tree node matches a lexical entry describing a vague temporal adverbial. The adverbial, together with the corresponding event, which is typically located in the parent node, are given to the `vaguetemp` property as arguments. The third argument is then the variable for the event instances that are filtered based on this information. In the following, this variable is part of the DUDES composition process, ensuring it is correctly integrated with the meaning and constraints of the other parts of the question.

The `vaguetemp` property is the interface between the NeoDUDES pipeline and `FuzzyLLI`, triggering the vague temporal adverbial evaluation as well as the generation of corresponding `FILTER` statements and bearing all necessary information for this process. More precisely, this is realized by adding a module to the SPARQL generation which reacts to the previously-generated special property `vaguetemp`. The module then calls (extended) `FuzzyLLI`, forwarding the event and vague temporal adverbial in order to get the temporal interval within which the event instance(s) referred to possibly happened. Based on this interval determined by `FuzzyLLI`, we can define the truth condition of the `vaguetemp` property as follows:

`vaguetemp(adverbial` adv`, event` ev`, event_instance` w`)`

$$\Leftrightarrow ref\_time - t_{start}^{adv,ev} \leq end(w) \leq ref\_time - t_{end}^{adv,ev}$$

where:

- ref_time denotes the reference time
- $end(w)$ is the time point at which the `event_instance` w ends
- $t_{start}^{adv,ev} = \text{argmin}_t(P_{\text{Adv}}(P_{\text{Ev}}(t)) \geq \theta)$
- $t_{end}^{adv,ev} = \text{argmax}_t(P_{\text{Adv}}(P_{\text{Ev}}(t)) \geq \theta)$, and
- θ denotes the chosen threshold for the possible temporal interval

This truth condition is then expressed in the form of SPARQL statements as follows:

```
?w ex:happensAt ?wInt . ?wInt time:hasEnd ?wEnd .
?w rdf:type ex:Eating .
```

`FILTER(xsd:dateTime(?wEnd) <=` $ref\_time - t_{end}^{adv,ev}$ `)`
`FILTER(xsd:dateTime(?wEnd) >=` $ref\_time - t_{start}^{adv,ev}$ `)`

If $t_{start}^{adv,ev}$ is ∞, the last line is omitted as it is always true. Moreover, in practice, the two `FILTER` statements are combined to a single one using `&&`. An illustration of these steps is presented in Figure 1, illustrating the DUDES representations for the different expressions in the question *"What did Tom eat a long time ago?"*. Note in particular that there is a DUDES for the expression "long time ago" that is combined with an event of type *"eating"*, and how the resulting interval is reflected in the corresponding SPARQL query. If no event is mentioned in the question, such as in *"What did Tom do recently?"*, or the event is not recognized as one of the supported events, the SPARQL query consists of a union of all possible event types and their associated intervals. In addition, just some minor adjustments for different parts of the pipeline to support the new knowledge graph and question types were necessary.

4 EXPERIMENTS

4.1 Dataset and Resource Construction

4.1.1 Dataset

For evaluation purposes, we rely on datasets from the WSU CASAS smart home project (Cook et al., 2013). The whole project contains 89 publicly available datasets[2], which vary by annotation status, number of participants, recording periods, seasonal contexts, among other factors. For our evaluation, we selected the dataset titled *twor.2010*, which includes sensor data from two participants living their daily lives in a smart home from August 23, 2009, to May 1, 2010. The sensor data are annotated by (Cook and Schmitter-Edgecombe, 2009), leading to the following thirteen events: *Bathing*, *Bed Toilet Transition*, *Eating*, *Enter Home*, *Housekeeping*, *Leave Home*, *Meal Preparation*, *Personal Hygiene*, *Sleep*, *Sleeping Not in Bed*, *Wandering in Room*, *Watch TV*, and *Work*. Each event annotation also specifies the participant (denoted as R1 or R2) performing the action.

4.1.2 Knowledge Graph and Evaluation Dataset

To construct the KG and evaluation dataset, we cleaned the data by first merging consecutive identical events and combining the similar labels "Sleeping Not in Bed" and "Sleep". We then removed "Sleep" and "Work" events shorter than 10 minutes and discarded "Wandering in Room", and finally merged consecutive identical events again.

The cleaned *twor.2010* dataset was used to build the KG, which is illustrated in Figure 2 using the exemplary event instance *Bathing_100* performed by *Tom*. The KG is structured as an RDF graph using the DUL and OWL-Time ontologies. Specifically:

- Pseudonyms were assigned to residents: "Tom" for R1 and "Mary" for R2.
- Key entities represented in the graph include:
  - **Agents:** Residents such as "Tom" and "Mary", instantiated as `rdf:type` of `dul:Agent`.
  - **Events:** Each of the eleven events is modeled as a `rdfs:subClassOf` of `ex:Event`.
  - **Time Intervals:** Each event instance happens at a temporal interval, modeled using `time:ProperInterval`. The start and end times of the interval are defined using `time:hasBeginning` and `time:hasEnd`.

The evaluation dataset consists of questions from four categories. Each question is associated with a reference time point, either randomly sampled within the dataset's overall time frame (August 23, 2009, to May 1, 2010), or deliberately set outside this range such that no event matches the question. Ground truth (GT) answers were determined using an extended version of `FuzzyLLI` (see Section 3.2 for the base model and Section 4.3 for the extension), in combination with our KG, with specific criteria defined per question category. We describe the question categories in the following:

For **"Did"** questions ("Did <resident_name> <event> <adverbial>?"), the GT is "Yes" if a corresponding event instance exists in the KG that is within the possible temporal interval of <event> and <adverbial> (determined by the extended `FuzzyLLI`) and was performed by <resident_name>, "No" otherwise. An equal distribution of "Yes" and "No" answers was achieved by adjusting the reference time.

"What" questions ("What has <resident_name> done <adverbial>?") are answered by extracting all event instances from the KG that match <resident_name>. Afterwards, we determine for each event instance and <adverbial> the possible temporal interval by using the extended `FuzzyLLI` model. The event is added to the GT if it happened within this possible temporal interval.

"What happened" For such types of questions, i.e. ("What happened <adverbial>?"), the question criteria are defined solely by the possible temporal interval determined by the <adverbial> and an event. Accordingly, we perform the same steps as for the

[2]Accessible via https://casas.wsu.edu/datasets/

Figure 2: Example KG, containing only the event instance *Bathing_100* performed by *Tom* at *Interval_Bathing_100* which is defined by a start and end time point in ISO 8601 format.

"What" questions but for all event instances regardless of the <resident_name>.

"Who" questions ("Who <event> <adverbial>?"). In this case, the criteria are defined by the possible temporal interval determined by using the extended `FuzzyLLI` model with <event> and <adverbial>. If the happening time of an event instance from <event> in the KG lies inside this temporal interval, the corresponding resident name is added to the GT. To generate cases with empty GTs, we also selected reference times from before the dataset's start date.

All <adverbial> values were among the four supported by `FuzzyLLI`: *just*, *recently*, *some time ago*, and *a long time ago*. In the case of *just*, the <adverbial> stands after the <resident_name> or after the question word for "What happened" or "Who". The eleven events were mapped to natural language phrases, e.g., "Bed Toilet Transition" $\rightarrow$ "go to the toilet", "Personal Hygiene" $\rightarrow$ "take care of personal hygiene", etc., used for <event>. In total, we automatically generated 2,780 questions, distributed as follows: 780 "Did", 800 "What", 400 "What happened", and 800 "Who" questions.

4.2 Query Selection

For query selection, we further fine-tuned the best-performing query selection model from (Schmidt et al., 2025) based on Flan-T5 (Chung et al., 2024) with a dataset based on all candidate queries generated by the pipeline for the above dataset. To underline adaptability to small amounts of data and account for the low linguistic diversity of the questions, we split the 2,780 questions into 20% train, 10% validation and 70% test splits.

The list of candidate queries was then slightly cleaned such that an F_1 score of 1 (i.e., perfect match) was only assigned to queries that contain at least one `FILTER` statement as well as `UNION` statements if the corresponding question is a "What" question.

The training dataset was created as described by Schmidt et al. (Schmidt et al., 2025), generating up to 100 comparisons per question. A hyperparameter search was performed comprising 20 trials using Optuna (Akiba et al., 2019), choosing a learning rate between 10^{-5} and 10^{-4} (logarithmic scale), a λ value for the lambda learning rate scheduler between 0.9 and 1.0 (logarithmic scale) as well as between 1 and 5 epochs. The best-performing model w.r.t. validation loss was chosen for the final evaluation. We used the same single-model strategies as proposed by (Schmidt et al., 2025) together with the upper-bound *BestScore* strategy that simulates a perfect query selection. For evaluation, the first (up to) 64 candidate queries were considered for each question.

4.3 Extension of `FuzzyLLI`

As outlined in Section 3.2, the original `FuzzyLLI` can not be generalized to unseen events, as each event has a specific standard deviation σ_e, used by the event-specific function (see Equation 1). This design ties the original model to events which are empirically measured with surveys like the one from (Kenneweg et al., 2024). To address this limitation and enable generalization across all eleven events in our KG, we propose an extension to the model: According to (Kenneweg et al., 2025a), based on the work of (Van Jaarsveld and Schreuder, 1985), each event is characterized by a *characteristic temporal signature*, defined by its typical duration and frequency. The parameters of the original `FuzzyLLI`, estimated in (Kenneweg et al., 2025a) support this hypothesis: *brushing teeth*, for instance, is characterized by a short duration and high frequency, resulting to a low standard deviation $\sigma_e = 935$, whereas *year abroad* has a long duration, low frequency, and consequently a high standard deviation $\sigma_e = 1{,}240{,}803$.

Accordingly, to generalize `FuzzyLLI` to our eleven events, we follow this hypothesis: A direct comparison between our events and those from the original `FuzzyLLI` is not feasible, as all of our events have a daily frequency and durations typically measured in minutes or hours. Among the original events (Brushing Teeth, Birthday, Vacation, Sabbatical, Year Abroad and Marriage), only *brushing teeth* shared a similar *characteristic temporal signature*. Since the original empirical data from (Kenneweg et al., 2024) did not include participants' expectations w.r.t. event duration and frequency, we used data from (Kenneweg et al., 2025b), who performed an extended version of the survey for seven events: They also asked participants to estimate both the typical duration and

frequency of each event.

An initial version of this extension used open-ended input fields for time units (seconds, minutes, hours, etc.), but the results were inconsistent and unreliable. Consequently, they followed a Likert-scale approach, which participants found easier to understand and complete. The scale for *duration* was: Minutes, Hours, Days, Weeks, Months, Years, Decades; and for *frequency*: Daily, Monthly, Yearly, Decadal, Once in a Lifetime.

We fitted this survey results to `FuzzyLLI` to determine σ_e for each of their events. Additionally, we estimated the typical duration of each event by taking the median of all survey responses. We used these data – the pair of (σ_e, duration) for each event – to train a (simplified) decision tree regression model. The *characteristic temporal signature* of an event is here only defined by its duration as the frequency of all our eleven events is "Daily". The resulting tree had depth 1, assigning a σ_e of $7,619$ minutes for events categorized as "minutes" and $22,367$ for events categorized as "hours".

Based on this decision rule, we manually categorized our events into those typically lasting minutes (*Bathing*, *Bed Toilet Transition*, *Eating*, *Enter Home*, *Meal Preparation*, *Personal Hygiene*) and those lasting hours (*Housekeeping*, *Sleep*, *Watch TV*, *Work*, *Leave Home*). This categorization leads to σ_e for each event as shown in Table 1.

Table 1: The duration-based σ_e (in minutes) of all our events. The duration is set manually by us.

Event	**Duration**	σ_e **(Minutes)**
Bathing Bed Toilet Transit. Eating Enter Home Meal Preparation Personal Hygiene	Minutes	$7,619$
Housekeeping Sleep Watch TV Work Leave Home	Hours	$22,367$

In conclusion, when the vague temporal adverbial evaluation is triggered in the extended pipeline via the property `vaguetemp` (see Section 3.3), the extended `FuzzyLLI` is provided with both the vague temporal adverbial and the corresponding event, and returns the "possible temporal interval" during which the adverbial applies above a defined threshold θ to the event. For our experiments, we set θ to 0.6, meaning the adverbial is considered to apply to at least a degree of 0.6 to the event during this "possible temporal interval". This interval is defined by start and end times relative to the reference time and is computed by the extended `FuzzyLLI` using the stored parameters μ_a and σ_a of the provided adverbial for the adverbial-specific function (see Equation 2), and the stored duration-based value σ_e of the provided event for the event-specific function (see Equation 1). For example, given the event *eat* and the adverbial *just*, the model returns an interval such as "0 to 140 minutes ago.". Similarly, for *eat* and the adverbial *long time ago*, it returns "$62,545$ to ∞ minutes ago."

5 RESULTS & DISCUSSION

The evaluation results of the extended NeoDUDES pipeline are presented in Table 2. The table contains two kinds of results. First, *BestScore* represents the best achievable score based on the generated candidate queries, thus indicating whether the queries generated by our approach are generally correct. Second, the table shows the results for the best-performing query selection model from the hyperparameter search combined with different strategies from (Schmidt et al., 2025) (*Accum* and *MostWins*, either accumulating the raw model outputs or counting separate wins for the final decision, respectively).

By design, the dataset contains questions and corresponding reference times for which the answer is empty, making micro or macro F_1 score evaluation impractical. Therefore, we present the number of exact matches in relation to the total number of questions, i.e., the exact match rate, in Table 2.

First, we can observe that the pipeline in principle generates correct queries among the candidate queries for all questions and reference times, as *BestScore* is 1.00. Similarly, the best-performing query selection model shows promising results for all tested query selection strategies, achieving exact match rates between 0.85 and 0.91.

Although these scores are very promising, they rely on a number of assumptions and preconditions. Most importantly, the pipeline's ability to translate vague temporal expressions into corresponding time intervals is limited by the events and temporal adverbials that `FuzzyLLI` can process. Events that do not occur daily, do not typically last minutes or hours, or lack a manually-specified duration in `FuzzyLLI`, as well as adverbials beyond the four supported ones can not be interpreted with the current pipeline and would demand further data and experiments. Moreover, we adapted the pipeline to support these types of questions, such that the success on a limited set of

questions is not particularly surprising. Similarly, the pipeline relies on lexical entries for those adverbials, events and all other relevant terms. Finally, the dataset is synthetic and limited w.r.t. question and SPARQL query diversity. An example SPARQL query for the question "Who ate a long time ago?" and reference date 2010-03-13T18:05:35.069542 is:

```
SELECT DISTINCT ?v3 WHERE {
  ?v1 dul:hasAgent ?v3 .
  ?v1 ex:happensAt ?v1Interval .
  ?v1Interval time:hasEnd ?v1End .
  ?v1 rdf:type ex:Eating .
  FILTER (xsd:dateTime(?v1End)
  <= xsd:dateTime("2010-01-29T07:40:35.069542")) }
```

The result of this query would be ex:Mary and ex:Tom. As we can see, the vague temporal adverbial "a long time ago" is transformed into a `FILTER` statement w.r.t. the corresponding event `ex:Eating` and reference time. Notably, "a long time ago" is interpreted as an interval with no lower bound, thus resulting in only one comparison included in the `FILTER` clause. As the question asks for the agent of those events, the query returns `?v3`, i.e., the object of the `dul:hasAgent` property.

"Did" questions asking for the existence of a corresponding triple pattern are thus very similar. An example for "Did Tom eat a long time ago?" including the intermediate steps is illustrated in Figure 1.

As we can see, the queries are very similar except for the reference time, some additional type checks and the query type being an `ASK` query. The structure of queries for questions of type "What happened" differs considerably from those, as all possible event types need to be considered in the query, leading to a complex disjunction. In the following, we see parts of an example for "What happened some time ago?":

```
SELECT DISTINCT ?v0 WHERE {
  ?v0 dul:hasAgent ?v3 .
  ?v0 ex:happensAt ?v0Interval .
  {
    ?v0Interval time:hasEnd ?v0End .
    ?v0 rdf:type ex:Bathing .
    FILTER ((xsd:dateTime(?v0End)
    <= xsd:dateTime("2010-01-24T09:16:23"))
    && (xsd:dateTime(?v0End)
    >= xsd:dateTime("2010-01-18T09:14:23")))
  } UNION {
    ?v0Interval time:hasEnd ?v0End .
    ?v0 rdf:type ex:Housekeeping .
    FILTER ((xsd:dateTime(?v0End)
    <= xsd:dateTime("2010-01-16T18:57:23"))
    && (xsd:dateTime(?v0End)
    >= xsd:dateTime("2009-12-30T04:08:23")))
  } UNION...
```

In contrast to questions with a fixed event like `ex:Eating`, "What happened" questions contain UNIONs of possible intervals for the respective adverbial, each constrained by the corresponding event. In such cases with no specific event mentioned in the question, the intervals for all known events have to be listed, yielding a long SPARQL query.

Table 2: Results for full dataset with best-performing (in terms of validation loss) query selection model.

Strategy	Exact Match Rate
Accum_{logits}	**0.91**
$\text{Accum}_{sigmoid}$	**0.91**
$\text{MostWins}_{0.0}$	0.85
$\text{MostWins}_{0.1}$	0.85
$\text{MostWins}_{0.25}$	0.86
$\text{MostWins}_{0.5}$	0.87
$\text{MostWins}_{0.75}$	0.88
$\text{MostWins}_{0.9}$	0.89
BestScore	1.00

All in all, our NeoDUDES pipeline extension illustrates the feasibility of including vague temporal expressions in QALD. Moreover, this shows the benefits of a modular and compositional question answering pipeline, which can therefore be easily adapted to support additional aspects of natural language and even for new domain-specific knowledge graphs without the need to manually create large amounts of training data.

6 CONCLUSIONS & FUTURE WORK

In this paper, we have presented an extension of the QALD system by (Schmidt et al., 2025) towards supporting questions with vague temporal adverbials. The interpretation of vague temporal adverbials in relation to a specific event relies on the extended `FuzzyLLI` model, a factorized probabilistic adverbial interpretation model introduced by (Kenneweg et al., 2025a). Our pipeline yields promising results, with exact match rates between 0.85 and 0.91 for the best-performing query selection model. Considering all generated candidate queries, our pipeline even generates correct queries for every question in the dataset. However, those scores have to be interpreted w.r.t. the limitations of our approach. For example, only four vague temporal adverbials (*just*, *recently*, *some time ago*, *long time ago*) and events that occur daily and have a duration in the range of "minutes" or "hours" are supported by the current implementation of the system. Further, the lexicon and some parts of the pipeline need to be extended for each new event to be supported by the system. Yet, the simplicity of the

pipeline extension shows the benefits of a modular, compositional QALD approach. Considering the limited question diversity of the evaluation dataset, future work could investigate other question categories such as "How often . . . ?" as well as involving Allen's relations (Allen and Ferguson, 1997), i.e., relations between two events, e.g., "Did Tom brush his teeth just before he ate?". In addition, besides events, the context in which a person speaks (prior communication, time, location) may also influence the interpretation of vague temporal adverbials. Finally, the query scoring model could be adapted to the temporal setting by including reference times in the model input.

ACKNOWLEDGEMENTS

This work is partially funded by the Ministry of Culture and Science of the State of North Rhine-Westphalia under grant no NW21-059A (SAIL) and by the Honda Research Institute Europe.

REFERENCES

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: a next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining*.

Allen, J. F. and Ferguson, G. (1997). *Actions and Events in Interval Temporal Logic*, pages 205–245. Springer Netherlands, Dordrecht.

Bobillo, F., Delgado, M., and Gómez-Romero, J. (2012). DeLorean: A reasoner for fuzzy OWL 2. *Expert Systems with Applications*, 39(1):258–272.

Bobillo, F. and Straccia, U. (2016). The fuzzy ontology reasoner *fuzzyDL*. *Knowledge-Based Systems*, 95:12–34.

Cavar, D., Dickson, B., Aljubailan, A., and Kim, S. (2021). Temporal Information and Event Markup Language: TIE-ML Markup Process and Schema Version 1.0. arXiv:2109.13892 [cs].

Chang, A. X. and Manning, C. (2012). SUTime: A library for recognizing and normalizing time expressions. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3735–3740, Istanbul, Turkey. European Language Resources Association (ELRA).

Chen, Z., Zhao, X., Liao, J., Li, X., and Kanoulas, E. (2022). Temporal knowledge graph question answering via subgraph reasoning. *Knowledge-Based Systems*, 251:109134.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Narang, S., Mishra, G., Yu, A., Zhao, V. Y., Huang, Y., Dai, A. M., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Cimiano, P. (2009). Flexible semantic composition with DUDES (short paper). In Bunt, H., Petukhova, V., and Wubben, S., editors, *Proceedings of the eight international conference on computational semantics, IWCS 2009, tilburg, the netherlands, january 7-9, 2009*, pages 272–276. Association for Computational Linguistics.

Cimiano, P., Unger, C., and McCrae, J. P. (2014). *Ontology-based interpretation of natural language*. Synthesis lectures on human language technologies. Morgan & Claypool Publishers.

Cook, D. and Schmitter-Edgecombe, M. (2009). Assessing the quality of activities in a smart environment. *Methods of information in medicine*, 48:480–5.

Cook, D. J., Crandall, A. S., Thomas, B. L., and Krishnan, N. C. (2013). CASAS: A Smart Home in a Box. *Computer*, 46(7):62–69.

Damerau, F. J. (1977). On "fuzzy" adjectives. *Linguistics*, 15(196):57–64.

de Moura, L. and Bjørner, N. (2008). Z3: An efficient SMT solver. In Ramakrishnan, C. R. and Rehof, J., editors, *Tools and algorithms for the construction and analysis of systems*, pages 337–340, Berlin, Heidelberg. Springer Berlin Heidelberg.

Frege, G. (1953). *The Foundations of Arithmetic: A Logico-Mathematical Enquiry into the Concept of Number*. Blackwell, Oxford.

Huang, R., Wei, W., Qu, X., Xie, W., Mao, X., and Chen, D. (2024). Joint Multi-Facts Reasoning Network For Complex Temporal Question Answering Over Knowledge Graph. arXiv:2401.02212 [cs].

Jia, Z., Pramanik, S., Saha Roy, R., and Weikum, G. (2021). Complex Temporal Question Answering on Knowledge Graphs. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, pages 792–802, New York, NY, USA. Association for Computing Machinery.

Jiao, S., Zhu, Z., Wu, W., Zuo, Z., Qi, J., Wang, W., Zhang, G., and Liu, P. (2023). An improving reasoning network for complex question answering over temporal knowledge graphs. *Applied Intelligence*, 53(7):8195–8208.

Kamp, H. and Sassoon, G. W. (2016). Vagueness. In Aloni, M. and Dekker, P., editors, *The Cambridge Handbook of Formal Semantics*. Cambridge University Press.

Kannen, N., Sharma, U., Neelam, S., Khandelwal, D., Ikbal, S., Karanam, H., and Subramaniam, L. (2023). Best of Both Worlds: Towards Improving Temporal Knowledge Base Question Answering via Targeted Fact Extraction. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on*

Empirical Methods in Natural Language Processing, pages 4729–4744, Singapore. Association for Computational Linguistics.

Kazakov, Y., Krötzsch, M., and Simančík, F. (2014). The Incredible ELK. *Journal of Automated Reasoning*, 53(1):1–61.

Kenneweg, S., Deigmoeller, J., Eggert, J., and Cimiano, P. (2025a). A factorized probabilistic model of the semantics of vague temporal adverbials relative to different events. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47.

Kenneweg, S., Deigmöller, J., Cimiano, P., and Eggert, J. (2025b). TRAVELER: A Benchmark for Evaluating Temporal Reasoning across Vague, Implicit and Explicit References. arXiv:2505.01325 [cs].

Kenneweg, S., Jackson, B. B., Deigmoeller, J., Eggert, J., and Cimiano, P. (2024). An Empirical Study on Vague Deictic Temporal Adverbials. In Zock, M., Chersoni, E., Hsu, Y.-Y., and de Deyne, S., editors, *Proceedings of the Workshop on Cognitive Aspects of the Lexicon @ LREC-COLING 2024*, pages 26–31, Torino, Italia. ELRA and ICCL.

Käfer, T. and Harth, A. (2018). Specifying, Monitoring, and Executing Workflows in Linked Data Environments. In *The Semantic Web – ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part I*, pages 424–440, Berlin, Heidelberg. Springer-Verlag.

Lange, L., Strötgen, J., Adel, H., and Klakow, D. (2023). Multilingual Normalization of Temporal Expressions with Masked Language Models. arXiv:2205.10399 [cs].

Mavromatis, C., Subramanyam, P. L., Ioannidis, V. N., Adeshina, A., Howard, P. R., Grinberg, T., Hakim, N., and Karypis, G. (2022). TempoQR: Temporal Question Reasoning over Knowledge Graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(5):5825–5833. Number: 5.

May, U., Zaczynska, K., Moreno-Schneider, J., and Rehm, G. (2021). Extraction and Normalization of Vague Time Expressions in German. In Evang, K., Kallmeyer, L., Osswald, R., Waszczuk, J., and Zesch, T., editors, *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 114–126, Düsseldorf, Germany. KONVENS 2021 Organizers.

McCrae, J. P., Spohr, D., and Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. In *Proceedings of the 8th extended semantic web conference on The semantic web: research and applications (ESWC)*, volume 6643, pages 245–259.

Neelam, S., Sharma, U., Karanam, H., Ikbal, S., Kapanipathi, P., Abdelaziz, I., Mihindukulasooriya, N., Lee, Y.-S., Srivastava, S., Pendus, C., Dana, S., Garg, D., Fokoue, A., Bhargav, G. P. S., Khandelwal, D., Ravishankar, S., Gurajada, S., Chang, M., Uceda-Sosa, R., Roukos, S., Gray, A., Riegel, G. L., Luus, F., and Subramaniam, L. V. (2021). SYGMA: System for Generalizable Modular Question Answering OverKnowledge Bases. arXiv:2109.13430 [cs].

Nenov, Y., Piro, R., Motik, B., Horrocks, I., Wu, Z., and Banerjee, J. (2015). RDFox: A Highly-Scalable RDF Store. In Arenas, M., Corcho, O., Simperl, E., Strohmaier, M., d'Aquin, M., Srinivas, K., Groth, P., Dumontier, M., Heflin, J., Thirunarayan, K., and Staab, S., editors, *The Semantic Web - ISWC 2015*, pages 3–20, Cham. Springer International Publishing.

Pustejovsky, J. (2005). Time and the semantic Web. In *12th International Symposium on Temporal Representation and Reasoning (TIME'05)*, pages 5–8. ISSN: 2332-6468.

Pustejovsky, J., Lee, K., Bunt, H., and Romary, L. (2010). ISO-TimeML: An International Standard for Semantic Annotation. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Schilder, F. and Habel, C. (2001). From temporal expressions to temporal information: semantic tagging of news messages. In *Proceedings of the workshop on Temporal and spatial information processing -*, volume 13, pages 1–8, Not Known. Association for Computational Linguistics.

Schmidt, D. M., Elahi, M. F., and Cimiano, P. (2025). Lexicalization Is All You Need: Examining the Impact of Lexical Knowledge in a Compositional QALD System. In Alam, M., Rospocher, M., van Erp, M., Hollink, L., and Gesese, G. A., editors, *Knowledge Engineering and Knowledge Management*, pages 102–122, Cham. Springer Nature Switzerland.

Sharma, A., Saxena, A., Gupta, C., Kazemi, S. M., Talukdar, P., and Chakrabarti, S. (2023). TwiRGCN: Temporally Weighted Graph Convolution for Question Answering over Temporal Knowledge Graphs. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2049–2060. arXiv:2210.06281 [cs].

Solt, S. and Gotzner, N. (2012). Experimenting with degree. In *Semantics and Linguistic Theory*, volume 22.

Stoilos, G., Stamou, G., Tzouvaras, V., Pan, J. Z., and Horrocks, I. (2005). Fuzzy OWL: Uncertainty and the Semantic Web. In *OWL: Experiences and Directions*.

Strötgen, J. and Gertz, M. (2010). HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions. In Erk, K. and Strapparava, C., editors, *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324, Uppsala, Sweden. Association for Computational Linguistics.

Su, M., Li, Z., Chen, Z., Bai, L., Jin, X., and Guo, J. (2024). Temporal Knowledge Graph Question Answering: A Survey. arXiv:2406.14191 [cs].

Van Jaarsveld, H. and Schreuder, R. (1985). Implicit quantification of temporal adverbials. *Journal of Semantics*, 4(4):327–339.