Utrecht University

Master Thesis

Monotonicity properties of *before* and *until*. An experimental study

Author:
Jan Winkowski

Supervisor:
Dr. Rick Nouwen

A thesis submitted in fulfilment of the requirements for the degree of Master of Science in Artificial Intelligence

in the

Utrecht Institute for linguistics OTS
Department of Languages, Literature and Communication

August 15, 2015

UTRECHT UNIVERSITY

Abstract

Faculty of Humanities
Department of Languages, Literature and Communication

Master of Science in Artificial Intelligence

Monotonicity properties of *before* and *until*. An experimental study

by Jan WINKOWSKI

The thesis concerns monotonicity properties of two temporal operators, namely *before* and *until*. Every natural language operator can be either monotone increasing, monotone decreasing or non-monotone. Usually, the first two mean that it preserves or reverses the direction of entailment. However monotonicity does not have to be defined in terms of entailment. In the paper three types of monotonicity are defined: with respect to temporal precedence, to the relation of being a sub-event, and to entailment. In this framework different definitions for *before* and *until* are proposed. These definitions allow to state exact empirical predictions. We test these predictions in an experimental study.

Acknowledgements

This thesis would be much worse (or would not be finished at all) without help and comments provided to me generously by many colleagues and friends. I would like to thank them here. Agnieszka Niemczyk and Manana Jaworska provided many helpful comments on the design of the experiment and data analysis. Katarzyna Kuś and Bartosz Maćkiewicz also commented on the design and helped to choose the proper polish expressions. Katarzyna Marzec drew a beautiful illustration of the temporal model. Marcin Bocian, Justyna Brzoska, Agata Niewiarowska and Małgorzata Winkowska helped gather the participants for the experiment and were happy to talk about the thesis whenever I wanted to. Yaron McNabb and Dominique Blok were kind enough to listen to my presentation. Finally, Rick Nouwen was the best supervisor I could have ever imagined. He was always there to talk about anything related to the thesis, accepted my method of working and did not care too much about the mistakes I made. Without him I would not even start the thesis, let alone finishing. I am also grateful to anyone who participated in the experiment, my friends and all the other people who provided help of any kind and that I have forgotten to mention.

Contents

List of Figures

List of Tables

Convention

The following notational convention will be used throughout the paper:

- $a, b, c, d, e, \ldots$ for event variables
- $p, q, r, s, \ldots$ for propositional variables
- $t_1, t_2, t_3, t_4, \ldots$ for timepoints
- $\pi_1, \pi_2, \pi_3, \pi_4, \ldots$ for periods
- $A, B, C, D, \ldots$ for sets
- $P, Q, R, S, \ldots$ for predicates of arbitrary arity
- $\mathcal{A}, \mathcal{T}, \mathcal{F}, \mathcal{G}, \ldots$ for functions

Indices will be skipped whenever they won't be needed. Any deviations from this convention will be explicitly stated. The arrow "$\Rightarrow$" is used for entailment, "$\rightarrow$" for material implication.

Let (A, B) where: A - is a non-empty set (domain), B - is a set of n-ary relations, be a relational structure. I.e. a structure in which there is a set of some entities (called a domain) and a set of relations defined over the domain. Relation R in a relational structure (A, B) is:

asymmetric iff $\forall a, b \in A[R(a, b) \rightarrow \neg R(b, a)]$

antisymmetric iff $\forall a, b \in A[(R(a, b) \wedge R(b, a)) \rightarrow b = a]$

intransitive iff $\forall a, b, c \in A[(R(a, b) \wedge R(b, c)) \rightarrow \neg R(a, c)]$[1]

[1]This property is sometimes called *antitransitivity*.

Chapter 1

Introduction

When I say "I have a red umbrella" it follows that I have an umbrella. But somehow, saying "I do not have a red umbrella" does not entail that I do not possess any umbrella at all. For some, this might be peculiar enough, but it gets worse when we consider a sentence: "I don't have an umbrella". At this point, beside being fed up with umbrella sentences, the reader might be sure that I do not posses a red umbrella.

We can infer all these sentences from the other sentences, because there is a particular connection between information conveyed in language and reasoning. On the semantic level this connection is called *entailment*. Since entailment is one of the most prominent mechanism in language inferences, it is important to understand how and why the aforementioned inferences are valid. These phenomena can be analysed by considering monotonicity properties of natural language inferences. At least since Ladusaw [1980] and Fauconnier [1975] it is believed that these properties can tell us a lot about an interesting class of natural language operators.

Probably most natural language operators are *monotone*.[1] This means that they are either *monotone increasing* (*upward entailing (UE)*), or *monotone decreasing* (*downward entailing (DE)*). The former usually means that they preserve entailment. Inferences licensed by these operators are inferences from subsets to supersets. More generally they preserve structure of some sort. Formally: let $\mathcal{Q}$ be an arbitrary operator, and let p, q be propositions. Then, $\mathcal{Q}$ is UE iff:

$$\forall p, q[(p \rightsquigarrow_1 q) \rightarrow (\mathcal{Q}(p) \rightsquigarrow_2 \mathcal{Q}(q))]$$

[1] Compare [Barwise and Cooper, 1981, p. 187]. In Universal 6 they propose: "The simple NP's of any natural language express monotone quantifiers or conjunctions of monotone quantifiers."

Where $\rightsquigarrow_{1,2}$ are given relations. In natural language analysis the interesting case is when $\rightsquigarrow_2$ is entailment. The $\rightsquigarrow_1$ can be defined in various ways, as will be done below. E.g. a determiner *some* is monotone increasing in its right argument because:

(1.1) Some elephants walk fast. $\Rightarrow$ Some elephants walk.

On the other hand, *monotone decreasing* or *downward entailing (DE)* operators reverse the order of entailment, or more generally, reverse given structure. Inferences triggered by these operators are inferences from supersets to subsets. Formally: let $\mathcal{P}$ be an arbitrary operator, and let p, q be propositions. Then, $\mathcal{P}$ is DE iff:

$$\forall p, q[(p \rightsquigarrow_1 q) \rightarrow (\mathcal{P}(q) \rightsquigarrow_2 \mathcal{P}(p))]$$

Where $\rightsquigarrow_{1,2}$ as above.

Negation is well known for reversing the order of entailment. E.g. a determiner *no* is monotone decreasing because:

(1.2) No elephants walk. $\Rightarrow$ No elephants walk fast.

An important of feature of Ladusaw's work, was the idea that, particular linguistic items (so called negative polarity items, NPIs for short) could only be licensed in DE contexts. A classic example is the expression *to lift a finger*. It can be properly used only within a negative context as in 1.3, but not in a positive context as in 1.4:

(1.3) She didn't lift a finger to help him.

(1.4) *She lifted a finger to help him.

Ladusaw claimed that polarity licensing depends on a derived logical meaning of a sentence. This is interesting, because it gives a clear example of a phenomenon primarily embedded in the semantics, and as such gives a strong argument why do we need semantic representation.[2] To sum up this short introduction into the concept of monotonicity, we should also remark that there exist functors which are neither UE, nor DE. These are called *non-monotone*.

The crucial thing to note here is that humans make inferences based on monotonicity automatically, without any difficulties, usually barely thinking about it. This is not the

[2] See Israel [2004] for a critical review.

case with computers. As presented below, a lot of effort is put into so-called entailment recognition, but this rarely concerns monotonicity properties.

In 2004 a generic task called *Textual Entailment Recognition (RTE)* was proposed (Dagan and Glickman [2004]). The definition of the task is reminiscent of entailment but only to some extent.

Definition 1.1. *Textual entailment* is defined as a directional relationship between pairs of text expressions, denoted by T (the entailing "Text") and H (the entailed "Hypothesis"). We say that T entails H if human reading would typically infer that H is most likely true. [Dagan et al., 2013, p. 3]

The task was meant to establish a framework which could unify most of the research done in connection with semantic inferences (such as: question answering, information retrieval, summarising, etc.). Until now, there were seven RTE-challenges, and the amount itself should be enough to show that the task is not easy. In the last challenge to date, the main task was to recognise textual entailment within a given corpus. Averages of the best system were: precision = 46.96 %, recall = 49.08 %, F-measure = 48 %.[3] Due to the lack of space, and the complicated nature of the task, reader is welcomed to see Bentivogli et al. [2011] for details.

Systematic understanding of entailment reversal could be very helpful in such tasks as mentioned above. For example, many of the systems employed in the RTE field use databases such as *WordNet*.[4] *WordNet* marks many semantic relations between the items it contains. Even though researchers try to exploit these relations in entailment recognition, sometimes it is not possible without paying attention to the monotonicity of the inference. E.g. *hyponymy* (*cat* is a hyponym of *pet*) preserves entailment in monotone increasing contexts, but not in the monotone decreasing ones. Compare:

(1.5) I have a cat. $\Rightarrow$ I have a pet.

(1.6) I don't have a cat. $\not\Rightarrow$ I don't have a pet.

This can lead to errors, since, as can be read in Dagan et al., p. 126:

> [...] robust recognition of the monotonicity direction still seems beyond state of the art in most applied systems. Thus, in practice, systems generally assume the upward monotonicity case, which is by far more frequent.

[3] Precision is the proportion of the retrieved cases that are relevant, recall is the fraction of relevant instances that are retrieved, and F-measure is a harmonic mean of these two.

[4] See Miller [1995] for an overview.

The aforementioned shows that recognition of monotonicity properties would be very welcomed in the RTE area. Some try to solve this problem by actually using Ladusaw's idea. In Danescu-Niculescu-Mizil and Lee [2010], Danescu-Niculescu-Mizil et al. [2009] the authors try to employ the connection between NPI's and DE contexts to construct an algorithm which could retrieve DE operators. Although the idea is definitely worth mentioning, the results are a bit controversial. E.g. quite a few items found by the authors does not necessarily make the related inference monotone decreasing. For example, take the finding which they mark as an "interesting case" i.e. *essential for*. The justification from the paper is [Danescu-Niculescu-Mizil et al., 2009, p. 143]:

(1.7) Talent is essential for singing. $\Rightarrow$ Talent is essential for singing a ballad.

However, note:

(1.8) (a) singing loudly $\Rightarrow$ singing

(b) Talent is essential for singing. $\nRightarrow$ Talent is essential for singing loudly.

The inference in 1.8 (b) does not hold since, even if talent is essential for singing, it is definitely not essential for singing loudly. There are more cases like the one above, but critique of the paper is not our goal. Especially since the authors seem to be aware of the problems, and their approach is definitely a step in the right direction.

In this paper I will not try to solve problems with monotonicity from the RTE field. However, such problems should be a convincing evidence that it is important to acquire knowledge about monotonicity patterns in natural language. This will be the goal of the work presented below. I will investigate entailment properties of two temporal operators, namely *before* and *until*. Formal investigations of such inferential patterns should give certain empirical predictions. These predictions will be tested in an experimental study. We shall now advance to different motivations for the presented research, namely: time and events.

1.1 Time and Events

An important aspect of how we experience reality is connected to the perception of time. Time itself is fundamental in almost every aspect of reality, and yet we are unable to investigate its nature directly. It might be that to some extent our experience of time is encoded in the language we use.

In Kamp [1979] H. Kamp discusses a particular way of representing time in discourse using events. As he points out, many of the information exchanged by people, in any meaningful conversation, concerns certain events. Usually the existence of the given event is presupposed by the participants, and the temporal relationship is given by the discourse itself. Since discourse participants treat those entities and the relations between them as primitives, maybe the easiest way of reconstructing such a discourse, is to start exactly with those primitives.

We shall also note that there are psychological and philosophical arguments for choosing events as basic building blocks in modelling time. We don't experience time directly, but rather we perceive it through events which have certain duration, or through change. We definitely can not perceive instants. As Bertrand Russell puts it in one of his lectures:

> Even if there be a physical world such as the mathematical theory of motion presupposes, impressions on our sense-organs produce sensations which are not merely and strictly instantaneous, and therefore the objects of sense of which we are immediately conscious are not strictly instantaneous. Instants, therefore, are not among the data of experience and must be either inferred or constructed. It is difficult to see how they can be validly inferred; thus we are left with the alternative that they must be constructed. [Russell, 1914, lecture IV] as quoted in Hamm and Bott [2014].

Such a construction can be done with events as primitives as will be presented below. The fact that we can construct instants out of events seem to be a good reason to include the latter in our ontology. But can we argue that they are more than just theoretical construction based on intuition?

Recently psychologists have started to examine whether events play a role in how we structure changing state of affairs. E.g. Zacks et al. [2001] found that their subjects perceived time in a pre-recorded movie, as consisting of discrete events. Even more, some of the findings suggest that the subjects were organising events into a hierarchical structure. I.e. each of the coarse grained events contained more finely grained events, and the events from a given level followed other events from this level. All this should be enough to justify the use of events in modelling time. In the next section we will discuss natural language means used to talk about time.

1.2 Temporal Prepositions

Temporal prepositions trigger fascinating inferential patterns and, simultaneously, they involve temporal reference. Since these are expressions, which we use in everyday language to talk about events in time, it seems that investigating them, could add something to our understanding of time perception.

Operators encoded in temporal prepositions seem to posses inference properties similar to quantifiers. For example (1.9) *before* is transitive as is *all*, and (1.10) *while* is symmetric similarly to *some*.[5,6]

(1.9)
$$\frac{\begin{array}{c}\text{A before B}\\ \text{B before C}\end{array}}{\text{A before C}} \qquad \frac{\begin{array}{c}\text{all A are B}\\ \text{all B are C}\end{array}}{\text{all A are C}}$$

(1.10)
$$\frac{\text{A while B}}{\text{B while A}} \qquad \frac{\text{some A are B}}{\text{some B are A}}$$

Recently, Zwarts and Winter [2000] made a successful attempt of employing a generalised quantifier approach toward prepositional phrases (PPs) with spatial prepositions. This seems also to be a right choice for studying PPs with temporal prepositions. In the aforementioned paper, the authors begin with the observation that functions denoted by spatial prepositions "should return entities with measurable distance and direction" (p. 4). This is motivated by realising that PPs can be modified by expressions which convey some additional information about distance or direction. Some examples follow (taken from Zwarts and Winter, p. 3):

(1.11) (a) The tree is behind the house. (bare PP)

(b) The tree is *ten meters* [behind the house]. (modified PP)

(1.12) *deep* under the castle, *diagonally* above the door, *far* outside the city, *right* in front of the car

As the authors explain:

---

[5] The inference with *while* in (2) might not be necessarily true, because of the contextual dependencies of *while*. The background event might be seen as the longer one. E.g. *She had lunch while playing games.* might not mean the same as: *She played games while having lunch.* However, when we treat it as an abstract operator it should hold.

[6] The examples, the way of presenting them, and many other things were inspired by Zwarts and Winter [2000].

> these structures are classified as PP modification [...] because the additional expression syntactically applies to a PP (or a P-bar) to produce another PP (P-bar). (p. 3)

In terms of modification, temporal prepositions bear certain resemblance to spatial prepositions. However it seems that in speech, time has only one dimension. If we depict this dimension as a line, there are only two variables which value could be changed: direction and length (amount). In reality, only one of these is modified. Consider these examples:

(1.13) shortly after dancing, 30 minutes before going out, almost until he was asleep.

In all of the above cases, the only effect of the modifier, is to refine the exact amount of time between the event in the main clause (omitted above), and the event in the temporal clause. Temporal direction of interest is already given by the preposition used. Therefore, we can generalise, the only effect of a proper modification of temporal expression is to delimit the "field of (temporal) action".[7] It seems, there are no modifiers which would apply to the direction itself.

All these observations show that there are similarities between spatial and temporal prepositions and between temporal prepositions and quantified statements. Hopefully we could exploit these similarities in order to investigate the inferences triggered by temporal prepositions. This will be done using tools provided by the theory of generalised quantifiers. I will try to give such an analysis with respect toward two temporal prepositions, namely *until* and *before.* As will be shown in the next section, these prepositions trigger interesting inference patterns. Possibly there is a connection between our perception of time and relations which are encoded in temporal prepositions. Since these relations manifests themselves in the linguistic reasoning concerning time, we will investigate such reasoning. This will be done by, first modelling the operators behind these two prepositions using events, and then by testing how well such models predict natural language use.

1.3 Before and Until

This section will concentrate on showing different intuitions concerning *before* and *until*, and how these intuitions were accounted for by different researchers. I will try to

[7]This excludes such modifiers as e.g. *only, exactly* which have much more complex behaviour, probably uniform across different arguments.

demonstrate that these accounts and intuitions often contradict each other, and therefore coming to definite conclusions without experimental studies is very difficult, if not impossible.

Before

Let's begin by considering certain intuitions from natural language. By saying that one event is before the other, we certainly want to express that the first event was earlier than the second. If this was all, it would be enough to model before with "<", i.e. temporal precedence relation. This is clearly not correct. Consider a sentence:

(1.14) Alice was driving before calling Mary.

Since *before* means here only "precedes in time", and there are two events in this sentence, there are also two possible interpretations. Either

(a) Alice was driving, she stopped, and then she called Mary, or

(b) Alice started driving at some time t_1, and then at some time t_2 she was still driving and she called Mary.[8]

Let's deal with the case (a) in the first place. When expressing the fact that one event was before the other, we don't necessarily want to say that the first one finished before the second one started. Often we want to say that the first event begun before the second one (which is the case in (b)). Consider longer events, and a relation expressed in the sentence:

(1.15) John was in Australia before Kyle was.

Usually, in such situations, we do not expect John to be back from Australia only to make the sentence true. Of course these sentences may have the reading in which the events do not overlap, but surely it is not the only reading available.

Considering the (b) case, note that here we already interpret *before* as more than just the precedence relation. This is because we explicitly assumed the existence of two different times such that one event starts in the first one, and continues at least until the second one, and the other event starts at the second time. Nothing like this is in the relation itself. This should suffice to show that we need some other rendition of *before*.

[8] If this was the case, I condemn Alice's unlawful behaviour.

However, *before* should have certain properties usually connected with the relation of temporal precedence. E.g. *before* seems to be transitive and asymmetric (as first pointed out by Anscombe [1964]). More data from natural language can be obtained if we consider some broad notion of monotonicity (why "broad notion" is going to be clarified later on). We might notice that *before* clauses should be monotone decreasing. This is because apparently NPIs are acceptable in before clauses (1.16), and because it licenses inferences from supersets to subsets. E.g. if you know that *Suzy danced during a party*, then the inference in 1.17 appears to be valid.

(1.16) She caught the baby before Tom *moved a muscle*.

(1.17) Suzy was running before partying. $\Rightarrow$ Suzy was running before dancing.

Since the the interpretation of *before* we are discussing at the moment is clearly inspired by Anscombe, and because we need to account for the monotonicity properties discussed above, our definition of *before* should satisfy these conditions:

Intuitions 1.1 (*Before 0*)**.** a before b is true iff:

(a) there is a time t_1 at which a happens,

(b) for all times t_2 such that b happens at t_2, t_1 precedes t_2

Universal quantification in (b) ensures that the temporal clause will be monotone decreasing, and taken together, these conditions are very similar to how Anscombe's proposal is treated in Beaver and Condoravdi [2003].

On the other hand universal quantification over times is problematic as pointed out by Landman [1991]. This is because a *before* definition which would use such quantification in the temporal clause could be vacuously satisfied, i.e. a *before* b is predicted to be true even if there is no time b. This is exploited by Beaver and Condoravdi in, what they call *Ketchup problems*. Consider the sentence (K):

(K) David ate lots of ketchup before he made a clean sweep of all the gold medals in the Sydney Olympics.[9]

If we follow the conditions given in Intuitions 1.1, (K) is predicted to be true, even if David never went to the Olympics.

[9] This is the example 32 in Beaver and Condoravdi.

We shall investigate how serious this objection is. Note that one might not want to exclude certain situations already in the semantics. By saying that, one thing was before the other we only put them in a certain relationship, and we might never be sure when this second thing will be fulfilled. I.e. there always is a possibility that David will be granted with all the gold medals from the Sydney Olympics at some later point in his life, because of some strange series of events.

If this is not convincing, we might always argue that there are pragmatic features which come into play. The whole persuasive force of the *ketchup problem* comes from the fact that the reader presupposes that David actually won all the gold medals at the Sydney Olympics. This happens because *before* is a presupposition trigger. In order not to let such strange things occur, we might want to reject the sentence when the presuppositions of the involved statements are not satisfied. In such case (K) would be neither true, nor false, but undefined.

However, one might really want to ensure that, when saying a *before* b, there exists some time t_2 in which b happens. This can be done by explicitly encoding it in semantics. This is, more or less, the approach taken by Heinämäki [1972] (as presented in Beaver and Condoravdi). Intuitive conditions behind this approach differ:

Intuitions 1.2 (*Before 1*)**.** a before b is true iff:

(a) there is a time t_1 at which a happens,

(b) there is a time t_2 at which b happens,

(c) t_1 precedes t_2

These two approaches are incompatible with each other. Possibly, one can always find some arguments which would make the first or the second more convincing. Fortunately, as we will see, definitions build upon each of these approaches predict different inferential patterns. These predictions can be tested, and this will be the method which will be followed here. We shall now discuss the case of *until*.

Until

A nice result shown in Kamp [1968], is that two two-argument operations *since* and *until* (definitions omitted) are *functionally complete* with respect to continuous linear orders. This means that any first-order temporal relation, which can be expressed in continuous linear order, can be expressed using only combinations of these two operators. This might suggest that *until* expresses a more powerful, and possibly more complex relation than *before*. First we shall see whether natural language intuitions accord with such a statement.

When we say:

(1.18) Suzy was writing until three in the morning.

we want to express at least that there was some time in which Suzy started writing, and that for all the times between that time and three in the morning, she was writing. It is not clear, whether she was *really* writing for all the time (compare Nouwen [2008]). E.g. if she had a small break, to pour herself a glass of water, we probably would say that she still was writing. The other thing which is not obvious at all, is whether she finished at three, or did she continue to some later time. It is difficult to decide what is the status of the endpoint in *until* sentences. It might be either encoded in the semantics or rendered as a pragmatic feature.

In the latter case, it is an implicature of sorts, and therefore it could be canceled. We could try such a cancellation and see whether it is clearly contradictory. Suppose, someone utters sentence 1.18, and then continues, in the manner presented below:

(1.19) Suzy was writing until three in the morning. In fact, she finished just before breakfast.

No contradiction arises. However, one might argue for the opposite case by pointing out two facts. They both have to do with the continuity of action.

First, it is possible that, we understand the situation depicted in 1.19 as not being continuous. I.e. Suzy was writing until three, than she had a (somehow meaningful) break, and then she continued until breakfast. This would mean that the endpoint is encoded in the semantics of the expression, and apparent cancellation of the implicature is not a cancellation at all, but nevertheless it forces the hearer to accord the meaning of two sentences in a way which resolves the contradiction.

The second argument is inspired by the study of properties of generalised quantifiers.[10] It might be that *until* is intransitive. Consider an inference below:

(1.20)

Peter was watching a movie until he started eating.
Peter was eating until midnight.

?? Peter was watching a movie until midnight.

[10] As done by Van Benthem [1984].

Whether the conclusion holds is dubious. If this was the case that in all such inferences, the conclusion does not follow, this would strongly suggest that the end-point is semantically encoded. Of course, these are not very convincing arguments, and we shouldn't treat them as such. They just show that it is very difficult to decide what is a proper semantic reconstruction of *until*, when one's views are only backed up by intuitions.

To add even more uncertainty, we shall reflect upon such three inferences, which validity depends on monotonicity properties of *until*:

$$\frac{\text{John was writing until the power went out.}}{\text{?? John was writing until the power went out and the alarm rang.}} \tag{1.21}$$

$$\frac{\begin{array}{l}\textit{eating event} \text{ is a part of } \textit{dinner event}\\ \text{John was writing until he started eating.}\end{array}}{\text{?? John was writing until dinner.}} \tag{1.22}$$

$$\frac{\begin{array}{l}\textit{running} \text{ precedes } \textit{taking a shower}\\ \text{John was stretching until he started running.}\end{array}}{\text{?? John was stretching until he took a shower.}} \tag{1.23}$$

Whether one want to accept any of these inferences depends on what they believe to be the exact meaning of until. There are at least three different possibilities. These will be presented by stating intuitive conditions which different definitions of *until* might satisfy.

Intuitions 1.3 (*Until 0*)**.** a until b is true iff:

(a) there is a time t_1 at which a happens,

(b) there is a time t_3 at which b happens,

(c) for all such times t_2 which are in between t_1 and t_3, a also happens in t_2

The interpretation above is the minimal interpretation. It only preserves what could be called continuity, i.e. that the event a goes on without stopping for a certain time. Note that here there is no constraint on the precedence of events involved. According to these conditions, if b precedes a, a *until* b would be vacuously satisfied. This might seem strange. These conditions assume that most of the features of *until* are due to pragmatic mechanisms it triggers.

Intuitions 1.4 (*Until 1*)**.** a until b is true iff:

(a) there is a time t_1 at which a happens,

(b) there is a time t_3, at which b happens,

(c) t_3 succeeds t_1,

(d) for all such times t_2 which are in between t_1 and t_3, a also happens in t_2

The conditions in the Intuitions 1.4 are the ones behind the aforementioned definition given in Kamp [1968]. A similar definition is also used in De Swart [1996]. The only difference between the Intuitions 1.3 and 1.4 is the point (c) in the latter. Here, it is explicitly stated which event precedes which.

Intuitions 1.5 (*Until 2*)**.** a until b is true iff:

(a) there is a time t_1 at which a happens,

(b) there is a time t_3, at which b happens,

(c) a doesn't happen in t_3,

(d) t_3 succeeds t_1,

(e) for all such times t_2 which are in between t_1 and t_3, a also happens in t_2

The conditions given above change the previous ones in a very small, yet meaningful detail. This is the explicit end-point encoded in the semantics. These conditions make the respective definition of *until* the most constrained.

One can see differences among these three intuitive definitions as if they were on a scale. On the left side of a scale there are expressions which carry the most semantic meaning and which allows for the least pragmatic interpretations. In the middle pragmatic and semantic parts are in balance. On the far right, the semantic meaning is the least determined which in turn reinforces pragmatic power. On the scale *Until 0* would be to the right, *Until 1* in the middle, *Until 2* is semantically the strongest one, and would be on the far left. As will be shown in the next section, these different conditions result in definitions which license different inferences.

Chapter 2

Logic and Semantics

2.1 Event Structures

We will now consider event structures as proposed in Kamp [1979]. Let event structure be a triple: $\mathbb{E} = \langle E, \prec, \bigcirc \rangle$. Where: E is a set of events, $\prec$ is a time precedence relation, $\bigcirc$ is an event overlap relation. The two relations satisfy postulates Ax 1 - Ax 6.

Ax 1. $\forall a \forall b (a \prec b \rightarrow \neg b \prec a)$

Ax 2. $\forall a \forall b \forall c ((a \prec b \wedge b \prec c) \rightarrow a \prec c)$

The precedence relation is asymmetric and transitive.

Ax 3. $\forall a (a \bigcirc a)$

Ax 4. $\forall a \forall b (a \bigcirc b \rightarrow b \bigcirc a)$

The overlap relation is symmetric and reflexive.

Ax 5. $\forall a \forall b (a \prec b \rightarrow \neg (a \bigcirc b))$

This property is called *separation*. If an event a precedes event b they can not overlap. Note, that since overlap is symmetric we also have: $\neg (b \bigcirc a)$.

Ax 6. $\forall a \forall b \forall c \forall d ((a \prec b \wedge b \bigcirc c \wedge c \prec d) \rightarrow a \prec d)$

This one is called *transfer*. One could see it as a way of transferring precedence by overlap. If a precedes b, b overlaps c, and c in turns precedes some d, then overlap works as precedence. So a precedes d.

Ax* 7. $\forall a \forall b(a \prec b \vee a \bigcirc b \vee b \prec a)$

Overlap and precedence taken together are *total*. Every event either precedes one another, or overlaps one another.

Whether we want to accept Ax 7 is dubious. Kamp, (p. 382) gives several reasons why this is so. Concertning our purposes, the strongest is probably the observation that usually in discourse, information about temporal relationship between events is not complete, and any attempt at modelling events in discourse, should take this into account e.g. but assuming only a partial order.[1]

Any event can be a temporal subset of the other. This is obtained using the aforementioned relations. For any two events a, b we have:

$$\begin{aligned} a \sqsubseteq b \text{ iff } & \forall e(e \prec b \rightarrow e \prec a) \text{ and} \\ & \forall e(b \prec e \rightarrow a \prec e) \text{ and} \\ & \forall e(e \bigcirc a \rightarrow e \bigcirc b) \end{aligned}$$

Note that this relation is not antisymmetric. Two events can be temporally equal and yet they don't need to be the same event. I.e. it can be that $c \sqsubseteq d$ and $d \sqsubseteq c$, but $d \neq c$ (see *working example* below).

2.2 Time Points and Propositions

We want a way to locate events in time which would be somehow independent of the events themselves. To do that we will use *instants* (I will use the notions *instants* and *time points* interchangeably).[2] [Kamp, 1979, p. 378] observes that, a maximal subsets of E of pairwise overlapping events can be used as such. Formally:

Instants. t is an instant (time point) in $\mathbb{E}$ iff

1. $t \subseteq E$

[1] As the author points out, this incompleteness of the information conveyed is especially visible in eyewitness reports (footnote 7):

> The speaker tries to piece the events as best as he can. The various aspects of the case may come back to him more or less at random. Often he will be able to reconstruct the true temporal sequence of the occurrences he relates, and to communicate this additional information, only in the latter part of the discourse. Sometimes he won't arrive at a complete articulation of the temporal relations at all.

[2] There are at least two ways in which one can obtain instants. In order to give a better review I discuss another one in the Appendix A.

2. $\forall e_1, e_2 \in t(e_1 \bigcirc e_2)$
3. $\forall e_1 \in (E \setminus t)\exists e_2 \in t[\neg(e_1 \bigcirc e_2)]$

Consider a structure:

$$\mathbb{T} = \langle I, < \rangle \text{ where:}$$

- I is a set of (all) instants in $\mathbb{E}$
- $t_1 < t_2$ iff $\exists e_1 \in t_1 \exists e_2 \in t_2(e_1 \prec e_2)$

It is easy to verify that this structure is a strict partial order. It would be linear, if we were to accept Ax 7. We can use this construction to "measure" our events in an objective manner. An event goes on at time t_1 if it is an element of t_1. An example can be find below.

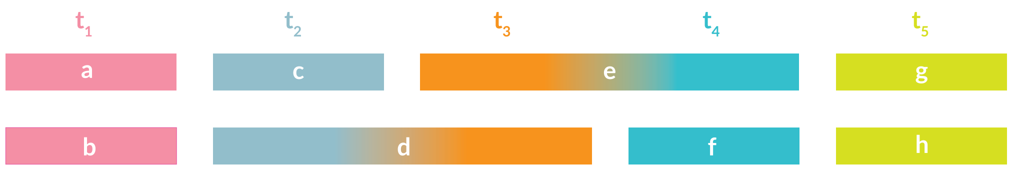

FIGURE 2.1: *A working example*

A working example[3]

Let $E = \{a, b, c, d, e, f, g, h\}$. Let relations between events in E be:

- Temporal precedence: $a \prec c, c \prec e, d \prec f, f \prec g$
- Subevent relation: $a \sqsubseteq b, b \sqsubseteq a, g \sqsubseteq h, h \sqsubseteq g$
- Overlap relation: $c \bigcirc d, d \bigcirc e, e \bigcirc f$

We can easily derive time points from this structure.
Time points: $t_1 = \{a, b\}, t_2 = \{c, d\}, t_3 = \{d, e\}, t_4 = \{e, f\}, t_5 = \{g, h\}$.

Note that: $t_1 < t_2 < t_3 < t_4 < t_5$. So even that, in the event structure, it is not the case that d precedes e, one can clearly see that d is (in some sense) earlier than e, because: $d \in t_3$ and $e \in t_4$, and $t_3 < t_4$. Compare Figure 2.1.

[3]This is inspired by the example from Landman, p. 194.

2.2.1 Propositions

We are interested in expressing certain relations between events. However, Kamp commits himself to a more ambitious task, i.e. to be able to model changes. We shall use his approach simplifying wherever possible.

Kamp defines two functions. Loosely speaking, first one returns the content of the event, the other one returns its state. We will only be concerned with the second one. Originally it maps atomic formulas of a language into four different values: $\mathbf{P}, \mathbf{F}, \mathbf{B}, \mathbf{C}$. We need only two of them in order to give the truth conditions in the model.

An event model is a tuple: $\mathbb{M} = \langle E, \prec, \bigcirc, \mathcal{A}_1, \mathcal{A}_2 \rangle$. Where $E, \prec, \bigcirc$ as before and $\mathcal{A}_1, \mathcal{A}_2$ are functions.

- $\mathcal{A}_2(e)(\varphi) = \mathbf{P}$ means: φ is true throughout e.
- $\mathcal{A}_2(e)(\varphi) = \mathbf{F}$ means: φ is false throughout e.

Assume we have an instant structure (a point structure described above) over such model: $T(\mathbb{M})$. We can define (simplified) truth in such a model.

$$[[\varphi]]_{T(\mathbb{M}),t} = 1 \text{ iff } \exists e \in t : \mathcal{A}_2(e)(\varphi) = \mathbf{P}$$

I will write $p(t_1)$ iff there is an event a such that, $a \in t_1$ and $\mathcal{A}_2(a)(p) = \mathbf{P}$. Formally:

$$p(t_1) \equiv_{\text{df}} \exists a[a \in t_1 \wedge \mathcal{A}_2(a)(p) = \mathbf{P}]$$

2.3 *Before* and *Until*

2.3.1 The definitions

Having the logical background settled, I shall present different definitions for *before* and *until*. These definitions are formalised versions of the *Intuitions* from chapter 1 (section 1.3). They will be put in terms of events. In 2.2.1 we've seen how to switch between events and propositions which are true during the respective events. However, since two out of three properties we are interested in, are stated in terms of temporal relations, it seems easier to give the definitions for events and switch to propositions only when necessary. I provide two examples of "propositional" variants for some of the definitions (marked with (b)). In a slightly informal way: if O is an operator on events, then it can be translated into its propositional variant O^* by substituting every expression of the form: $e \in t$ with the expression of the form: $p(t)$.

Before

Definition 2.1.

$$(a)\ BEF_0(a,b) \equiv_{\mathrm{df}} \exists t_1[a \in t_1 \wedge \forall t_2(b \in t_2 \rightarrow (t_1 < t_2))]$$

$$(b)\ BEF'_0(p,q) \equiv_{\mathrm{df}} \exists t_1[p(t_1) \wedge \forall t_2(q(t_2) \rightarrow (t_1 < t_2))]$$

Definition 2.2. $BEF_1(a,b) \equiv_{\mathrm{df}} \exists t_1[a \in t_1 \wedge \exists t_2((b \in t_2) \wedge (t_1 < t_2))]$

Until

Definition 2.3.

$$(a)\ UNT_0(a,b) \equiv_{\mathrm{df}} \exists t_1 \exists t_3[(a \in t_1) \wedge (b \in t_3) \wedge \forall t_2(t_1 < t_2 < t_3 \rightarrow a \in t_2)]$$

$$(b)\ UNT'_0(p,q) \equiv_{\mathrm{df}} \exists t_1 \exists t_3[p(t_1) \wedge q(t_3) \wedge \forall t_2(t_1 < t_2 < t_3 \rightarrow p(t_2))]$$

Definition 2.4.

$$UNT_1(a,b) \equiv_{\mathrm{df}} \exists t_1 \exists t_3 > t_1[(a \in t_1) \wedge (b \in t_3) \wedge \forall t_2(t_1 < t_2 < t_3 \rightarrow a \in t_2)]$$

Definition 2.5.

$$UNT_2(a,b) \equiv_{\mathrm{df}} \exists t_1 \exists t_3 > t_1[(a \in t_1) \wedge (b \in t_3) \wedge (a \notin t_3) \wedge \forall t_2(t_1 < t_2 < t_3 \rightarrow a \in t_2)]$$

2.3.2 Monotonicity

We can posit at least three natural types of monotonicity in the event structures. I will begin by defining these notions, and by giving some examples (these may be sometimes dubious). Then, I will prove monotonicity properties for each of the definitions.

Note that, from the relational point of view, temporal prepositions are of the form: $\mathcal{T}(A,B)$. So, as is the case with quantifiers, we can talk about "left" and "right" monotonicity. Below are the definitions for the "right" monotonicity. The other one is defined analogously. A function is non-monotone if it is neither monotone increasing nor monotone decreasing.

EMON and PMON are defined on events not on propositions. This might seem problematic, since nothing in natural language expresses events. But we already know how to obtain the corresponding propositions (using the functions $\mathcal{A}_1$ and $\mathcal{A}_2$), and these are expressed by sentences.

MON 1. Temporal function $\mathcal{T}$ is event monotone (EMON)

(a) increasing (EMON↑) iff $\forall e, a, b \in E[(a \sqsubseteq b) \rightarrow (\mathcal{T}(e, a) \rightarrow \mathcal{T}(e, b))]$

(b) decreasing (EMON↓) iff $\forall e, a, b \in E[(a \sqsubseteq b) \rightarrow (\mathcal{T}(e, b) \rightarrow \mathcal{T}(e, a))]$

A function is event monotone when it preserves or reverses the containment of events. In natural language there is no expression of such relation as event inclusion, but it is possible to give some approximating examples. We shall write e.g. assume washing hair ⊑ taking shower. This should be understood as: "assume that any event of hair washing is temporally included in an event of taking a shower" (which could be roughly translated to: whenever a person is taking a shower she is washing her hair). Examples:

Assume: dancing ⊑ partying
EMON↓ John took a shower before partying.
John took a shower before dancing.

Assume: washing hair ⊑ taking shower
↑EMON John washed his hair before going out.
John took a shower before going out.

MON 2. Temporal function $\mathcal{T}$ is precedence monotone (PMON)

(a) increasing (PMON↑) iff $\forall e, a, b \in E[(a \prec b) \rightarrow (\mathcal{T}(e, a) \rightarrow \mathcal{T}(e, b))]$

(b) decreasing (PMON↓) iff $\forall e, a, b \in E[(a \prec b) \rightarrow (\mathcal{T}(e, b) \rightarrow \mathcal{T}(e, a))]$

A temporal function may preserve or reverse temporal precedence. The former is fulfilled when from the fact, that one event is earlier than the other follows, that what happens in the first event entails what happens in the second. The latter is the case when the entailment is reversed. Assumptions in the examples mean that one event was earlier than the other. Examples:

Assume: dancing ≺ talking
PMON↑ Suzy was drinking before dancing.
Suzy was drinking before talking.

	Assume: cleaning floors $\prec$ closing doors
↓PMON	Suzy closed the doors before going home.
	Suzy cleaned the floors before going home.

MON 3. Temporal function $\mathcal{T}$ is entailment monotone (ENTAIL)

(a) increasing (ENTAIL↑) iff $\forall p, q, r[(p \Rightarrow q) \rightarrow (\mathcal{T}(r,p) \rightarrow \mathcal{T}(r,q))]$

(b) decreasing (ENTAIL↓) iff $\forall p, q, r[(p \Rightarrow q) \rightarrow (\mathcal{T}(r,q) \rightarrow \mathcal{T}(r,p))]$

Note, that above the function $\mathcal{T}$ is still over events, not propositions. This is what one could call "proper" monotonicity. A temporal function preserves entailment when it licenses inferences from subsets to supersets. It reverses entailment when it licenses inferences from supersets to subsets. Assumptions here are straight forward, but I list them anyway. Examples:

	Assume: writing furiously $\Rightarrow$ writing
↑ENTAIL	Peter was writing furiously until making a mistake.
	Peter was writing until making a mistake.

	Assume: power went out and the alarm rang $\Rightarrow$ power went out
ENTAIL↓	Anna was writing before the power went out.
	Anna was writing before the power went out and the alarm rang.

Below I will prove monotonicity properties for the proposed definitions.

2.3.3 Proofs

Before

Proposition 1. *BEF_0 is ↑EMON↓.*

Proof. Let $a, b, c, d, e \in E$.
For ↑EMON we need to show that, if $a \sqsubseteq b$, then $BEF(a,c) \rightarrow BEF(b,c)$.
Assume $a \sqsubseteq b$. Additionally assume $BEF(a,c)$. To show: $BEF(b,c)$.

From the definition of BEF : $BEF(a,c) \equiv \exists t_1[(a \in t_1) \wedge \forall t_2((c \in t_2) \rightarrow (t_1 < t_2))]$.
Take such a and t_1. Since $a \sqsubseteq b$, so from the definition of subevent, $\forall e(e \bigcirc a \rightarrow e \bigcirc b)$.
Therefore, if a is in some t, so is b. Particularly, $b \in t_1$.
Recall that t_1 is such that $BEF(a,c)$. Hence, $BEF(b,c)$.

For EMON$\downarrow$ we need to show that, if $b \sqsubseteq c$, then $BEF(a, c) \rightarrow BEF(a, b)$.
Assume: $b \sqsubseteq c$. Assume further: $BEF(a, c)$ i.e. $\exists t_1[(a \in t_1) \wedge \forall t_2((c \in t_2) \rightarrow (t_1 < t_2))]$.
To show: $BEF(a, b)$ i.e. $\exists t_1[(a \in t_1) \wedge \forall t_2((b \in t_2) \rightarrow (t_1 < t_2))]$.

Take a and t_1 as given above. Take arbitrary t such, that $b \in t$.
To show: $t_1 < t$ i.e. $\exists e \in t_1 \exists e' \in t(e \prec e')$.

Note that, because $BEF(a, c)$ is assumed, we have $\forall t_2((c \in t_2) \rightarrow (t_1 < t_2))$.
From the definition, $t_1 < t_2$ means that: $\exists e'' \in t_1 \exists e''' \in t_2(e'' \prec e''')$.
From the assumption that $b \sqsubseteq c$: $\forall d(d \prec c \rightarrow d \prec b)$.
In particular $e'' \prec c$, because $e'' \in t_1$ and $t_1 < t_2$. Hence, $e'' \prec b$.
We assumed earlier that $b \in t$. Therefore, $t_1 < t$. □

Proposition 2. *BEF_0 is $\downarrow$PMON$\uparrow$.*

Proof. Let $a, b, c \in E$.

For $\downarrow$PMON it suffices to show that: if $a \prec b$, then $BEF(b, c) \rightarrow BEF(a, c)$.
Assume: $a \prec b$ and $BEF(b, c)$. To show: $BEF(a, c)$.

From the definition of BEF we have: $\exists t_1[(b \in t_1) \wedge \forall t_2((c \in t_2) \rightarrow (t_1 < t_2))]$.
Note, we have Ax 5 (*separation*): $\forall a \forall b(a \prec b \rightarrow \neg(a \bigcirc b))$.
We assumed that $a \prec b$. From this and the definition of a time point $a \notin t_1$.
But, since $b \in t_1$, a must be in such t_0 that $t_0 < t_1$.
We have: $t_0 < t_1$ and $t_1 < t_2$. Hence, from the transitivity of "$<$": $t_0 < t_2$.
Therefore: $BEF(a, c)$.

Similarly for PMON$\uparrow$. □

Proposition 3. *BEF_0 is $\uparrow$ENTAIL$\downarrow$.*

Proof. We need to show, that for any model $\mathbb{M}$, for any two sentences p, q such that $p \Rightarrow q$, BEF preserves (or changes the direction of) entailment.

For $\uparrow$ENTAIL Assume: $BEF(p, r)$ and $[[p \Rightarrow q]]_{T(M)} = 1$. To show: $BEF(q, r)$.
$BEF(p, r)$ means: $\exists t_1[p(t_1) \wedge \forall t_2(r(t_2) \rightarrow (t_1 < t_2))]$.
Note that for all t if $p(t)$, then $q(t)$. In particular, since we have $p(t_1)$, also $q(t_1)$ holds.
Therefore $BEF(q, r)$.

For ENTAIL$\downarrow$ assume: (1) $BEF(p, r)$ and (2) $[[q \Rightarrow r]]_{T(M)} = 1$. To show: $BEF(p, q)$.
From (2), we have for all t: if $q(t)$, then $r(t)$.
Suppose $q(t')$. Then, $r(t')$.
But then, from (1), there is t_1 and $p(t_1)$ and, since $r(t')$, $t_1 < t'$.

Hence, for all such t' in which $q(t')$, there is t_1 such that $t_1 < t'$, and $p(t_1)$.
Therefore $BEF(p, q)$. □

Proposition 4. *BEF_1 is ↑EMON↑.*

Proof. For ↑EMON we need to show that if $a \sqsubseteq b$, then $BEF(a, c) \rightarrow BEF(b, c)$.
This is similar to the same case for BEF_0, i.e. Proposition 1.

For EMON↑ we need to show that if $b \sqsubseteq c$, then $BEF(a, b) \rightarrow BEF(a, c)$.
Assume $b \sqsubseteq c$, and $BEF(a, b)$. To show: $BEF(a, c)$.

$BEF(a, b)$ means: $\exists t_1[a \in t_1 \wedge \exists t_2((b \in t_2) \wedge (t_1 < t_2))]$.
Take such a and t_1. To show: $\exists t_2((c \in t_2) \wedge (t_1 < t_2))$.

From $b \sqsubset c$, and the definition of subevent, we have $\forall e(e \bigcirc b \rightarrow e \bigcirc c)$.
Therefore, if b is in some t, so is c. Particularly, $c \in t_2$.
From the assumption $t_1 < t_2$, therefore $BEF(a, c)$. □

Proposition 5. *BEF_1 is ↓PMON↑.*

Proof. For ↓PMON we have to show: if $a \prec b$, then $BEF(b, c) \rightarrow BEF(a, c)$.
Assume $a \prec b$ and $BEF(b, c)$. To show: $BEF(a, c)$.

From the definition, $BEF(b, c) \equiv \exists t_1[b \in t_1 \wedge \exists t_2((c \in t_2) \wedge (t_1 < t_2))]$.
Take t_2 and c as above. We have to show that there is t_1, in which a is, and $t_1 < t_2$.
This follows in the same manner as in the proof of Proposition 2.

For PMON↑ we have to show: if $b \prec c$, then $BEF(a, b) \rightarrow BEF(a, c)$.
Assume $b \prec c$ and $BEF(a, b)$. To show: $BEF(a, c)$.

From $BEF(a, b)$ we have: $\exists t_1[a \in t_1 \wedge \exists t_2((b \in t_2) \wedge (t_1 < t_2))]$.
Take such a and t_1. We have to show: $\exists t((c \in t) \wedge (t_1 < t))]$.
From the assumption that $b \prec c$, we can infer that c is in t such that $t_2 < t$.
Note that $<$ is transitive, so $t_1 < t$. Therefore, $BEF(a, c)$. □

Proposition 6. *BEF_1 is ↑ENTAIL↑.*

Proof. The proof is similar to the proof of Proposition 3. □

Until

Proposition 7. *UNT_1 is (-)PMON(-).*

Proof. For (-)PMON it is sufficient to show that, it is neither ↑PMON, nor ↓PMON. Counterexamples to each of these statements are shown below.

Take the structure described in section 2.2 (under *A Working Example*).
Let $E = \{a, b, c, d, e, f, g, h\}$ be set of events, such that:

- $a \prec c, c \prec e, d \prec f, f \prec g$
- $a \sqsubseteq b, b \sqsubseteq a, g \sqsubseteq h, h \sqsubseteq g$
- $c \bigcirc d, d \bigcirc e, e \bigcirc f$

We can generate points out of these events. Let $I = \{t_1, t_2, t_3, t_4, t_5\}$ be a set of points (where: $t_1 = \{a, b\}, t_2 = \{c, d\}, t_3 = \{d, e\}, t_4 = \{e, f\}, t_5 = \{g, h\}$). It is easy to see that: $t_1 < t_2 < t_3 < t_4 < t_5$.

For ↑PMON assume $a' \prec b'$ and $UNT_1(a', c')$. To show: $\neg UNT_1(b', c')$.
Let $a' = c, b' = e, c' = d$. Then, trivially: $c \prec e$. Note $c \in t_2$ and $d \in t_2$ and $d \in t_3$, and that there is no t in between t_2 and t_3. Therefore $UNT_1(c, d)$ is vacuously satisfied. However, $UNT_1(e, d)$ is not satisfied, because there is no such t which would contain e and would precede some t containing d.

For ↓PMON assume $a' \prec b'$ and $UNT_1(b', c')$. To show: $\neg UNT_1(a', c')$.
Let $a' = a, b' = e, c' = f$. Since $a \prec c, c \prec e$, and $\prec$ is transitive, also $a \prec e$.
Similarly as above, $UNT_1(e, f)$ holds. But $UNT_1(a, f)$ does not, because a is not in every t in between t_1 and some t containing f. E.g. it is not in t_2.

Similarly, for PMON(-) we have to show, that it is neither PMON↑ nor PMON↓. We are going to use the same structure as above.

For PMON↑ assume $b' \prec c'$ and $UNT_1(a', b')$. To show: $\neg UNT_1(a', c')$.
Let $a' = c, b' = d, c' = f$. Obviously, we have: $d \prec f$.
$UNT_1(c, d)$ follows, as shown before. However, $UNT_1(c, f)$ does not, because there is a time point which does not contain c, and is in between every time point that contain c or f. Namely, such time point is t_3.

For PMON↓ assume $b' \prec c'$ and $UNT_1(a', c')$. To show: $\neg UNT_1(a', b')$.
Let $a' = e, b' = d, c' = f$. Then $d \prec f$ as previously.

$UNT_1(e, f)$ is vacuously satisfied, similarly as in the counterexample for PMON↑.
The same counterexample explains why $UNT_1(e, d)$ is not satisfied. □

Proposition 8. *UNT_1 is ↑EMON↑.*

Proof. Let $a, b, c \in E$.
For ↑EMON assume: $a \sqsubseteq b$ and $UNT_1(a, c)$. To show: $UNT_1(b, c)$.

$UNT_1(a, c)$ i.e. $\exists t_1 \exists t_3 > t_1[(a \in t_1) \wedge (c \in t_3) \wedge \forall t_2(t_1 < t_2 < t_3 \rightarrow a \in t_2)]$.
Take such t_1 and t_3. Since $a \sqsubseteq b$, there is t such that, both $a, b \in t$.
Let this $t = t_1$. Hence. $UNT_1(b, c)$.

For EMON↑ assume: $b \sqsubseteq c$ and $UNT_1(a, b)$. To show: $UNT_1(a, c)$.

$UNT_1(a, b)$ i.e. $\exists t_1 \exists t_3 > t_1[(a \in t_1) \wedge (b \in t_3) \wedge \forall t_2(t_1 < t_2 < t_3 \rightarrow a \in t_2)]$.
Take such t_1 and t_3. Since $b \sqsubseteq c$, from the definition of subevent, $\forall e(e \bigcirc b \rightarrow e \bigcirc c)$.
Therefore, if b is in some t, so is c. Note that we assumed: $UNT_1(a, b)$.
Hence, $UNT_1(a, c)$. □

Proposition 9. *UNT_1 is ↑ENTAIL↑.*

Proof. Let $M = \langle E, <, \bigcirc, \mathcal{A}_1, \mathcal{A}_2 \rangle$ be a model of our event structure and $T(M)$ be instant structure over such model. Let $a, b \in E$ and let p, q be sentences.

For ↑ENTAIL assume $[[p \rightarrow q]]_{T(M)} = 1$, $UNT_1(p, r)$. To show: $UNT_1(q, r)$.

From $[[p \rightarrow q]]_{T(M)} = 1$ follows $[[p \rightarrow q]]_{T(M),t_1} = 1$.
And since $[[p]]_{T(M),t_1} = 1$, so $[[q]]_{T(M),t_1} = 1$. Hence, $UNT_1(q, r)$.

Similarly for ENTAIL↑. □

We shall now prove the properties of UNT_0 and UNT_2. Only some proofs will be shown, since most of them is similar to the respective proofs for UNT_1.

Proposition 10. *UNT_0 is ↑EMON↑.*

Proof. Similar to the proof of Proposition 8. □

Proposition 11. *UNT_0 is ↑PMON↓.*

Proof. For ↑PMON we have to show: if $a \prec b$, then $UNT_0(a, c) \rightarrow UNT_0(b, c)$.
Assume $a \prec b$ and $UNT_0(a, c)$. To show: $UNT_0(b, c)$.

From the definition, $UNT_0(a,c) \equiv \exists t_1 \exists t_3[(a \in t_1) \wedge (c \in t_3) \wedge \forall t_2(t_1 < t_2 < t_3 \rightarrow a \in t_2)]$.
Take t_3 and c as above. We have to show $\exists t(b \in t \wedge \forall t'(t < t' < t_3 \rightarrow b \in t')]$.
There are four possibilities. Either $c \prec b$, or relation between b and c is not given, or $b \bigcirc c$, or $b \prec c$.

If $c \prec b$, or the relation between them is not given, then there is no such t' which would be in between t and t_3, and $UNT(b,c)$ is vacuously satisfied.

If $b \bigcirc c$, then either $b \in t_3$ and $UNT(b,c)$ is also vacuously satisfied, or it is in some t''.
If $t'' < t_3$, then $UNT(b,c)$ holds for the reason given below.
If $t_3 < t''$, then $UNT(b,c)$ holds for the reason given above.

If $b \prec c$, then, from the assumption, it has to be between a and c.
In such case, recall that $UNT(a,c)$. Hence, also $UNT(b,c)$.

For PMON↓ we have to show: if $b \prec c$, then $UNT(a,c) \rightarrow UNT(a,b)$.
Assume $b \prec c$ and $UNT(a,c)$. To show: $UNT(a,b)$.
This is shown similarly as in the previous case. □

Proposition 12. *UNT_0 is ↑ENTAIL↑*

Proof. Similar to the proof of Proposition 9. □

Proposition 13. *UNT_2 is ↑EMON(-)*

Proof. ↑EMON follows in the same way as was presented in the proof of Proposition 8.

For EMON(-), we have to show that it is neither EMON↑, nor EMON↓.
As before, this will be done using the structure from the proof of Proposition 7.

For EMON↑ assume $a' \sqsubseteq b'$ and $UNT(c', a')$. To show: $\neg UNT(c', b')$.
Let $a' = f, b' = e$ and $c' = d$. It can be easily seen that $f \sqsubseteq e$.
$UNT(d, f)$ holds, but $UNT(d, e)$ does not, because $d \in t_3$ and $e \in t_3$.

For EMON↓ assume $a' \sqsubseteq b'$ and $UNT(c', b')$. To show: $\neg UNT(c', a')$.
Let $a' = f, b' = e$ and $c' = c$. Again $f \sqsubseteq e$.
It is easy to check that $UNT(c, e)$ is satisfied.
Note that $c \in t_2, f \in t_4$, and crucially $c \notin t_3$.
Hence, $UNT(c, f)$ is not satisfied, which concludes the proof. □

Proposition 14. *UNT_2 is (-)PMON(-)*

Proof. Similar to the proof of Proposition 7. □

Proposition 15. *UNT_2 is ↑ENTAIL↑*

Proof. Similar to the proof of Proposition 9. □

Chapter 3

Experimental Study

3.1 Introduction

In this part I will talk about an experiment I've conducted, in order to check which of the definitions render correctly the use of temporal operators in natural language. First I will state the motivations behind the experiment and what other work influenced the experiment. Then, I will move to the proper report of the experiment.

3.1.1 Motivation

As we have already seen there is no consensus among the linguists and the logicians what is the proper logical representation for the temporal operators *before* and *until*. Many (if not most) of the arguments against (or pro) some of these definitions are arguments appealing to the intuitions from natural language. This seems fair: since we want to account for all proper uses of some expression, the fact that a definition renders some of the seemingly correct uses incorrect (or some incorrect uses correct), is a strong argument against that definition.

However, linguists' intuitions can often be mistaken. There is a reason why natural language is called *natural*. If one spends most of the time investigating a certain phenomenon, this phenomenon stops being obvious and natural to her at some point. This is also the case with natural language, and this makes reasonable asking ordinary folk for their intuitions on certain natural language phenomena.

This might seem a good enough reason. Still, there is more to that than just checking to what extent the intuition of the laypeople accord with the intuitions of the linguists. At least since the beginning of the XXI century a growing body of experimental work appeared, that shows important differences between logical and contextualised interpretation of many natural language expressions.[1] This movement stems particularly from theoretical work of H.P. Grice (e.g. Grice [1989]) but also, to some extent, from the earlier works done in the philosophy of language which emphasised use of language in opposition to pure logical meaning (e.g. Austin [1962], Wittgenstein [1953]). It uses tools from psycholinguistics to show how people reason in language and how different cognitive capacities are employed in language use. This is closely connected to the research in such prominent areas as “embodied cognition”, “ecological rationality”, and similar.

The aforementioned, is another important motivation for the experiment described below. The monotonicity of an expression (see Chapter 1 and Section 2.3.2 from Chapter 2) is an important linguistic property, but to my current knowledge no one ever tested monotonicity of particular operators in an experimental setting.[2] From the cognitive point of view, the ability to recognise the proper inferential properties of an expression has important consequences. The subject might not be fully aware of such recognition but its effects are there. Below some of them will be presented.

Geurts [2003] and Geurts and van Der Slik [2005] are concerned with the effects monotonicity of quantifiers has on the difficulty of reasoning. In the second paper, beside their main task, they give couple arguments why monotonicity is important. For example, a nice argument which might show that monotonicity is relevant to the psychology of language production and comprehension is this: the fact that NPIs (see Chapter 1) occur only in monotone decreasing environments, is as much a fact about language, as it is a fact about speakers. Probably no competent speaker will ever utter a sentence containing e.g. a word *any* in an environment of a positive polarity. Geurts and van Der Slik, (p. 101) suggest that this is so, because the speakers “routinely compute the monotonicity properties of incoming and outgoing utterances.” Moreover, it seems that we learn such distributional constraints of NPIs quite early. Studies show (see O’Leary and Crain [1994] and Musolino et al. [2000] for an overview) that children by the age of four already refrain from using words such as *anything* in the contexts of positive polarity.

Another interesting phenomenon connected with monotonicity and reasoning is a particular property of preserving validity of inferences, even in cognitively vague environments. This has to do with so-called *donkey sentences*. Consider a sentence:

[1] For example: Noveck [2001], Bott and Noveck [2004], Huang and Snedeker [2009], Chemla [2008], Chevallier et al. [2008], and many others.

[2] However, there were some experiments which concerned monotonicity in general. See papers by Geurts, and Geurts and van Der Slik referenced in the same paragraph. The latter is especially worth reading. The following paragraphs are based on Section 3 from this paper.

(3.1) Every farmer who owns a donkey beats it.

It is not very clear how this should be interpreted. I.e. suppose there is a farmer who owns two donkeys but beats only one. Is 3.1 true in this situation? Or does every farmer is needed to beat every donkey they own, for this sentence to be true? Even though speakers intuitions tend to differ, surprisingly, as suggested by Kanazawa [1994] and Geurts [2002], speakers seem to agree on certain inferences involving such sentences. An example from Geurts and van Der Slik, (p. 102) is:

(3.2) Every farmer who owns a male donkey beats it. (entailed by 3.1)

(3.3) Every farmer who owns a donkey beats it with a stick. (entails 3.1)

The inferences above are preserved due to monotonicity properties of *every*. As the authors conclude (p. 103):

> Apparently, monotonicity inferences may go through even if the interpretation of a sentence remains underdetermined or unclear in certain respects.

Beside these well described effects of monotonicity properties, we might hypothesise that there are other aspects in which they influence the way people communicate. For example an utterance of an expression with a given monotonicity turns the hearer attention more toward certain aspects of the utterance than the other. Take the *umbrella* sentence from the very beginning of the paper (Chapter 1):

(3.4) I don't have an umbrella.

It seems pointless to ask, whether the speaker has a red umbrella, or an umbrella of any other colour. However, it would not be, if the speaker would utter a sentence:

(3.5) I have an umbrella.

This is because each of these sentences trigger different presuppositions, and allows the hearer to draw different immediate conclusions. Quite possibly this is connected to different monotonicity properties of each of these sentences.[3] If this is indeed so,

[3]Of course the sentence 3.4 is an explicit negation of the sentence 3.5. This makes our example difficult to generalise, because some of these phenomena might be connected closer to negation than to monotonicity. However, it is worth noting, that these two go hand in hand. For an excellent study of negation see Horn [1989].

then any study concerning monotonicity brings a lot into our understanding of natural language.

Given all this, the motivation behind the experiment should be clear. I want to check how certain temporal expressions function within the natural language setting and how well monotonicity properties account for natural language inferences. The experiment should also add something to the work on the distinction between semantics and pragmatics.

3.1.2 Inspiration, experimental paradigms and problems

The most straightforward way of testing monotonicity properties is to ask subjects to infer consequences from certain assumptions, using a given expression. There are many problems with such a task though.

Even simple deductive reasoning can be very demanding when put as an abstract problem.[4] This suggest, that such an experiment, has to have a carefully chosen, natural setting. Also, ideally, we would want the subjects to be careful about the use of a particular expression, and not necessarily about the reasoning itself. Therefore, we want to stress the use of the expression and make the reasoning as natural as possible. Finally, since we are testing the logical properties of certain expressions, we want to be sure that the subjects get the literal reading, not contaminated by any pragmatic features.

In order to show how one can overcome such problems, below I present short reviews of a few experiments the design of which was the most inspirational and helpful.

Bott and Chemla on *Free Choice*

In a series of beautifully designed experiments, E. Chemla and L. Bott investigated the so called *free choice* inferences (Chemla and Bott [2014]). These are inferences, in which a disjunctive sentence can be interpreted as a conjunction.

The experimental paradigm used, was a variation on the seminal paradigm devised by Bott and Noveck [2004]. The main idea behind it was that, for some items, every answer the subjects provided had exactly one interpretation. They used *underinformative* sentences such as:

(3.6) Some elephants are mammals.

[4] Perfect example is Wason's selection task (Wason [1968]).

Whether the subject decides to agree or disagree depends on whether they draw the pragmatic inference. *Some* carries a scalar implicature which could be interpreted by the participants as "some and not all". Therefore, a "no" answer means that the subject draws this inference and does not agree with the fact that there are elephants which are not mammals. Answering "yes" on the other hand, means that the subject interprets the sentence in harmony with its logical reading, and agrees with the fact that *some* does not exclude *all*.

Chemla and Bott changed the original setting in a way which allowed for the manipulation by the introduced context, and reduced the link to the background knowledge. The subjects were given a story in which earth was facing a global catastrophe. Certain people were allowed to save certain types of items. If a person was a zoologist she was allowed to save a living creature, if she was an engineer she was allowed to save an artificial object. Once again, target sentences were constructed in such a way, that answering "yes" could be connected to only one of the two possible readings (similarly with "no" answer). For example, when subject read a sentence:

(3.7) Derek-the-engineer is allowed to save a hammer or a lion.

and answered "yes", it meant that they interpreted the sentence in a literal way (engineers were allowed to save hammers). A "no" answer meant that the subject derived the *free choice* inference (since engineers were not allowed to save lions). It is worth noticing, how the change they provided makes the paradigm suitable for testing many different phenomena.

Cognitive load and pragmatic/semantic distinction

There is a substantial amount of work devoted to showing differences in processing the semantic part of an expression and its pragmatic counterpart. This was inspired by two competing theories on the derivation of implicature.[5]

Many of these experiments try to increase the cognitive load the subjects have to deal with, in order to show that one of the reading is cognitively easier. The trailblazers in the field were Bott and Noveck who have shown, that when the time to answer is limited, subjects tend to prefer logical reading over scalar implicature. This was replicated in many different experiments, e.g. De Neys and Schaeken [2007] used a dual task paradigm to find a similar result, Breheny et al. [2006] arrived at the same

[5]The *default view* and the *relevance theory*. E.g. Levinson [2000] and Sperber and Wilson [1986] respectively.

conclusion using a self-paced reading setting, etc. Apparently varying the cognitive load can help in distinguishing the pragmatic interpretation from the semantics.

3.2 Experiment

The experiment design was based on a paradigm devised by Chemla and Bott. Because of the particular, exploratory, character of the experiment, I wanted to distinguish between pragmatic and semantic inferences. This was done by adding a time limit.

To review the motivation for the experiment: what are the relations hidden behind the natural language use of *before* and *until*? What inferences are licensed by these operators?

3.2.1 Experimental setting

There is only a small set of possible monotonicity patterns for the investigated operators. This is because (as mentioned before), a function denoted by an operator usually has two arguments, and in each of the arguments it can be either monotone increasing, monotone decreasing or non-monotone. Therefore, there are at most six possibilities. By testing each case we can constructively decide which pattern is the most robust. So, a trivial hypothesis is: inferential patterns licensed by the operators *before* and *until* will be correctly described by one of the proposed definitions. Since our definitions describe all the most probable cases, this hypothesis shall turn out to be true. We may try to refine the hypothesis a bit. This can be done by recalling the already discussed accounts.

Beaver and Condoravdi [2003] are mostly concerned with the veridicality properties of *before*, so with respect to inferential properties, their definition closely resembles the one given by Anscombe [1964]. This points to the definition BEF_0. The definition given by Heinämäki [1972] is a bit different when it comes to monotonicity properties. In fact it suggest, the proper definition should be BEF_1. Mainly because *before* is a presupposition trigger Anscombe's resolution seems more convincing. Therefore, in case of *before* the refined hypothesis is: BEF_0 will give correct predictions with respect to monotonicity properties of the operator.

The case of *until* is much easier, since the widely accepted (e.g. De Swart [1996]) definition is the one given originally by Kamp in Kamp [1968]. This suggest that, *until* should behave in accordance with the definition UNT_1.

Variables and Materials

The experiment was conducted in Polish. *Before* was rendered using a word *przed*, and *until* using the construction *aż do*. This decision was discussed with Polish linguists but it was still arbitrary to some degree. In Polish there are at least three different ways of expressing *before* and *until*, so it had to be such. Hopefully the best possible choice was made.

Below an exemplary target item will be shown, and each variable will be described with the use of this item. An item consists of three entities:

E1 Timetable

E2 Assumption (“A sentence”)

E3 Consequence (“B sentence”)

A timetable consists of statements presenting quantified information about events in time within certain environment. The cover story for the experiment was “a day in the kindergarden”, so the timetable was an excerpt from the class schedule from a made-up kindergarden. An example follows.

- **10.00 - 11.00** Second Breakfast
- **11.00 - 14.00** Fun and games, including: **13.00 - 14.00** Games in groups
- **14.00 - 15.00** Lunch

An assumption is a sentence which provides some information about a fact concerning a child’s day in the kindergarden. It gives a name of a child, an activity, and what is a temporal relation between this activity and the timetable. The assumption uses the target operator. The target sentences were all of the form:

$$X \ [activity] \ [temporal\ operator] \ [timetable\ item]$$

where X is the name of the child with the indication of a group to which it belongs (it was only one word in Polish), an activity was a verb phrase, describing something a child could do during the day in the kindergarden (it was never an item from the timetable, see the list of activities in the appendix), a temporal operator was either the target operator (*before, until*) or a filler operator (these were *after*, *while*, or a sentence without an operator i.e. a control sentence), and a timetable item was an item from the timetable. Examples:

- Jacob - from the small group was drinking water before games in groups.
- Suzy - from the middle group was playing tag before music.

Finally, a consequence was a sentence having the same form as the assumption, but it differed in the temporal clause. What the difference was exactly depended on which inference the item was supposed to test. If it was PMON↑, the item from the timetable in the assumption, would be earlier in the timetable than the respective item in the consequence. If it was EMON↓, then the item in the assumption would temporally contain the item in the consequence (temporal containment was additionally marked on the timetable with the word "including"), and if it was ENTAIL↑, then the temporal clause in the assumption would entail the temporal clause in the consequence. The other types would differ in the direction of the respective inference.

For each item, the task of the participant was to judge whether the consequence is true or false in the light of the evidence given by the information from the timetable and the assumption. This presents most of the variables of which the experiment consisted. The independent variables were: the temporal operator, the monotonicity of the reasoning, and the time to answer. How the first two were manipulated should be obvious by now, the manipulation of the third one was performed by introducing two groups with different amount of time to answer (see below for details). This was done to introduce time pressure which could help to distinguish between the literal reading and the pragmatic reading, similarly as in previously discussed experiments. The dependent variable was the validity of the sentence.

Method

There were two experimental groups which differed in the amount of time the subjects had for providing the answer. In each group there were two values of the dependent variable, TRUE and FALSE. This was operationalised with a question ("Is the B sentence true?"), which had two possible answers ("Yes" and "No"). Most TRUE judgments for a given item would mean, that the item is of a given monotonicity, and most FALSE answers would mean that it is not, respectively.

All possible independent variable levels were: two target temporal operators, two filler temporal operators, three predicted monotonicity patterns for each target operator, and two time frames. In theory, each temporal operator has two arguments, and in each argument it can be either monotone increasing, decreasing, or non-monotone (see Table 3.1). If one would want to test each of these patterns, one would need four measurements per monotonicity pattern. Since there are three patterns, this gives twelve measurements for each operator.

Definition	EMON	PMON	ENTAIL
BEF_0	↑EMON↓	↓PMON↑	↑ENTAIL↓
BEF_1	↑EMON↑	↓PMON↑	↑ENTAIL↑
UNT_0	↑EMON↑	↑PMON↓	↑ENTAIL↑
UNT_1	↑EMON↑	(-)PMON(-)	↑ENTAIL↑
UNT_2	↑EMON(-)	(-)PMON(-)	↑ENTAIL↑

Table 3.1: Comparison of different monotonicity properties.

However, it can be easily noticed that, there is no need to test the left argument of any of the operators. This is because, such a test does not give us any information which could not be gathered by testing the right argument. So, we can cut the number of needed measurements in half. Then, again, one measurement for each of the cases does not seem enough. There is always a danger that, the chosen item has some inherent properties which could distort the results. Therefore additional measurement was introduced for each item. Since, there are two target operators, there were twenty four target items. It is considered good practice to hide target items within at least the same amount of filler items.

This was done by adding sixteen items which resembled closely the target items, but they were constructed using different temporal operators (the aforementioned *while* and *after*). The eight remaining items were control sentences. These were checking the basic comprehension of the timetable. The assumption gave a name of a child, a group to which it belonged, and what time it was. The consequence stated simple fact, which either was in accord with the information in the timetable, or was against it. By adding twenty four filler items, we arrive at the forty eight items per group, and this was the final amount of items in the experiment.

The test for monotonicity in natural language is, to check whether a given proposition entails certain other proposition. Given this, we interpret the truth judgments provided by the participants in the experiment, as if it was given (more or less) by one person. In order to facilitate the interpretation of the answers, the reader will find a table (Table 3.2) comparing all the possibilities below.

Participants and Procedure

In total seventy four participants took part in the experiment. They were volunteers, mainly gathered through various web-based social services. Some of them were passers-by found in the public libraries. The latter were rewarded with a small candy for taking

Definition	EMON↑	EMON↓	PMON↑	PMON↓	ENTAIL↑	ENTAIL↓
BEF_0	No	Yes	Yes	No	No	Yes
BEF_1	Yes	No	Yes	No	Yes	No
UNT_0	Yes	No	No	Yes	Yes	No
UNT_1	Yes	No	No	No	Yes	No
UNT_2	No	No	No	No	Yes	No

Table 3.2: Interpreting the results.

part in the experiment. None of the subjects were excluded from the final analysis (see below).

The experiment was conducted in Polish using a free internet service called IBEX. First a participant read the instructions and was asked to provide some basic personal information. Then there was a four item practice session. The procedure in the practice session was the same as in the proper experiment, but it concerned a different subject. Subjects were asked to judge the truth of sentences about the height of certain people.

Immediately before the items, the subject was presented with a long version of a timetable. In the practice session the table concerned height of a group of people. In the main experiment it was a timetable of a two groups from a kindergarden. The participant was asked to get acquainted with the table and proceed whenever she felt she was ready. Then the first item was shown.

In every item the subject was first shown the shortened version of the appropriate timetable and a short horizontal line indicating where the assumption will appear. The assumption (marked with "A") was presented after pressing the space-bar key. The sentence remained underlined until the subject pressed the space-bar again. Then, the underlined consequence (marked with "B") appeared beneath the assumption, and the assumption ceased to be underlined. Both sentences remained on screen for a specified amount of time, which was either 4000 ms (short lag) or 10000 ms (long lag). Then, the sentences and the timetable disappeared, and the question ("Is the sentence B true?") appeared. Under the question there were two, horizontally placed, words: "Yes" and "No". The subject could provide the answer either by clicking the word, or by a keystroke. After giving the answer, subject proceeded to another item.

During the whole time of the experiment a progress bar was shown on screen. Before the first and the second space-bar press, a sentence with different graphic properties was shown in the lower part of the screen, which reminded that, the subject should press the space-bar. Also, to remind the subjects what the task is, a question (written in a smaller typeface) was displayed beneath the timetable.

Subject were assigned to the different time groups in a random manner, and they were not aware to which group they were assigned, or even, that there were two groups. The order of the items was also random, but every target item was followed by a filler item (there was no distinction between a filler and a control, so there was a possibility of a sequence of eight target items with eight control items between them).

3.2.2 Results

Subjects

Seventy four participant were mainly women ($n = 54 = 73\%$) and only few men ($n = 20 = 17\%$). The oldest participant was sixty seven years old, the youngest was nineteen (mean = 31.8, median = 27). Most of the subjects had higher education ($n = 61 = 82\%$), the others finished high school ($n = 13 = 18\%$). Only one participant declared nationality other than polish (french), and two declared that their mother tongue was english or french. Usually subjects spoke at least two foreign languages (min = 0, max = 7, median = 2).

Two subjects who declared mother tongue other than polish were excluded from the data. No other subjects were excluded. This might seem wrong when compared with the results of control questions. There were two participants who answered four (= 50%) control questions incorrectly (i.e. not in accordance with the intention of the experimenter), nobody did more. However, a lot of participants answered at least some of the controls incorrectly ($n = 24 = 33\%$). This definitely shows that there was a certain problem with the control item (and maybe in general with the design of the experiment), but this also makes indicating a particular number, over which a subject should be excluded, difficult. Therefore, I decided to keep all the subjects despite their sometimes poor performance in the control questions. This will be discussed in detail in section 3.3.2.

Since there were much more women than men, I've tested how significant was the difference in answers with respect to the declared gender. This was done using a χ^2 test. The results: $p = 0.140, \chi^2 = 2.180$, show that the difference was not significant. However, the sample of the answers provided by men might be too small to be absolutely convinced.

Experimental Data

Thirty two out of seventy two subjects (=44%) were assigned to the long lag group. This resulted in 768 answers to target items in the long lag and 960 answers in the short lag. First, we tested whether there was a significant difference between the answers provided in each of the group per target operator. The results were quite surprising. In case of before there was no significant difference: $p = 0.634, \chi^2 = 0.226$. The answers for until, on the other hand, were significantly different: $p = 0.011, \chi^2 = 6.525$. The summarised results from both time groups are presented below.

	Before		**Until**	
	Yes %	No %	Yes %	No %
EMON↑	26	74	19	81
EMON↓	83	17	26	74
PMON↑	83	17	10	90
PMON↓	17	83	19	81
ENTAIL↑	82	18	78	22
ENTAIL↓	47	53	31	69

TABLE 3.3: Summarised results for *before* and *until*

Before

Even though no significant difference was found globally, the results in the long lag and the short lag were different. This is shown in the Table 3.4 (* marks value that exceeds 100% due to rounding).

	Long lag		**Short lag**	
	Yes %	No %	Yes %	No %
EMON↑	23	77	29	71
EMON↓	83	17	83	18*
PMON↑	81	19	84	16
PMON↓	16	84	19	81
ENTAIL↑	84	16	80	20
ENTAIL↓	45	55	49	51

TABLE 3.4: Differences for *before* in the short lag and the long lag

To be completely sure that there is no significant difference in neither of the conditions, we tested those that were different, for significance. The results show no significant differences. EMON↑ long lag vs short lag: $p = 0.472, \chi^2 = 0.517$, ENTAIL↑ long lag vs short lag: $p = 0.498, \chi^2 = 0.460$.

To facilitate the interpretation of the results, they are visualised in a bar graph in the Figure 3.1.

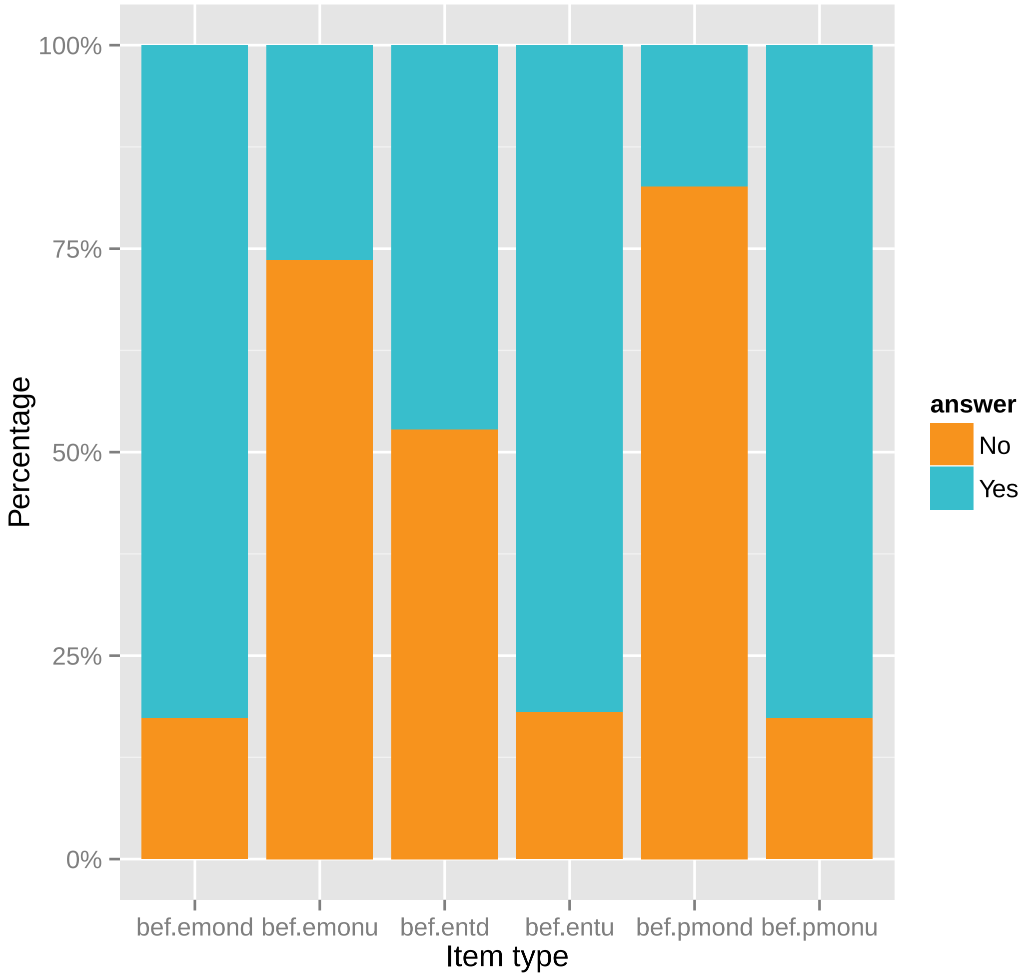

Figure 3.1: Summarised results for *before*

Until

In the case of *until* there were some important differences noted between the short lag and the long lag. These are compared in the Table 3.5.

	Long lag		Short lag	
	Yes %	No %	Yes %	No %
EMON↑	17	83	21	79
EMON↓	19	81	33	68*
PMON↑	2	98	16	84
PMON↓	19	81	20	80
ENTAIL↑	75	25	81	19
ENTAIL↓	27	73	35	65

TABLE 3.5: Differences for until in the short lag and the long lag

Once again we tested the significance of the differences within different conditions (presented from the most significant to the least significant).

type	*p - value*	χ^2
PMON↑	0.003	8.739
EMON↓	0.063	3.461
ENTAIL↓	0.278	1.178
ENTAIL↑	0.364	0.822
EMON↑	0.541	0.375
PMON↓	0.851	0.035

TABLE 3.6: Significance of the differences between the short lag and the long lag.

Since the changes between the long lag and the short lag were only positive (i.e. in every condition more participants were accepting the consequence in the short lag than in the long lag), the most effective way of visualising the changes between the time groups, would be to present the differences between the proportion of "Yes" answers in each condition. This is presented in the Figure 3.2.

3.3 Discussion

The discussion is divided into two parts. First I present the interpretation of the results for both prepositions. Than I discuss possible confounds.

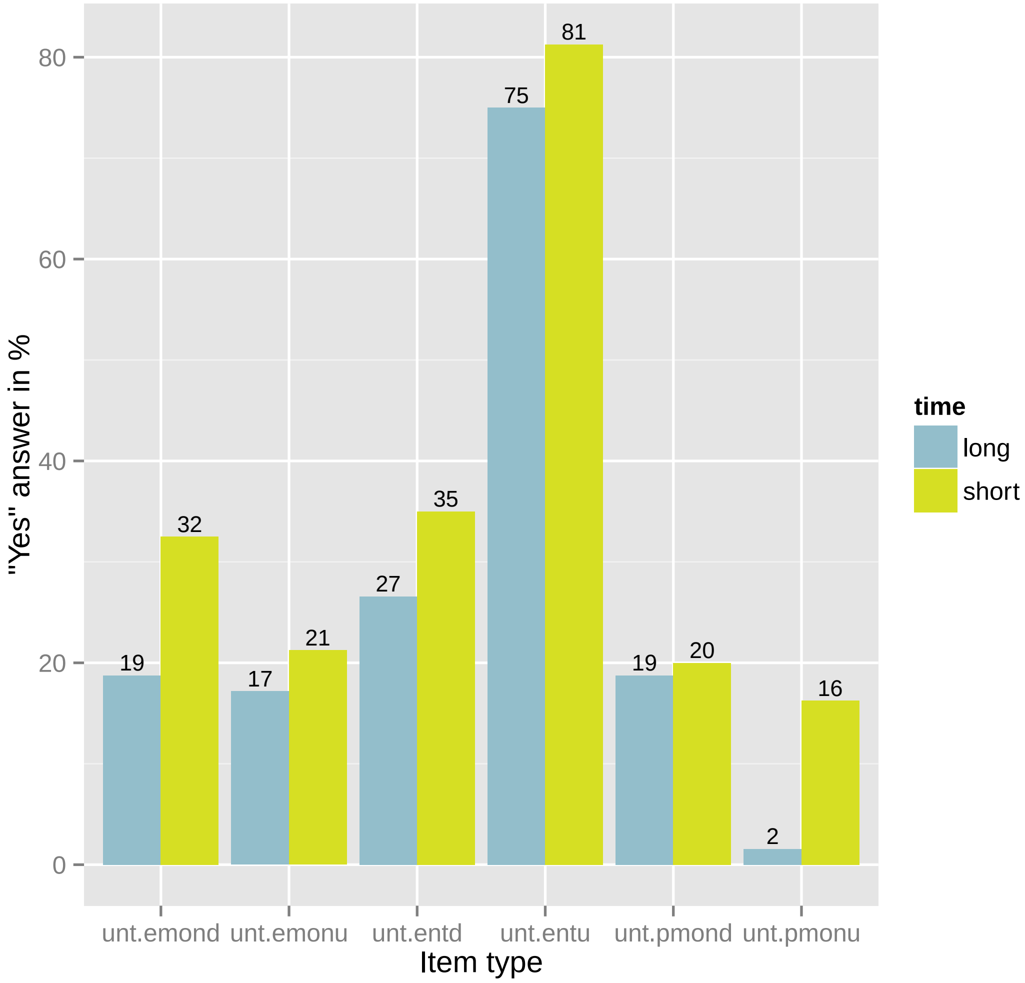

FIGURE 3.2: Change in the proportion of the positive answers for *until*.

3.3.1 Results interpretation

Before

The results for *before* are quite straightforward. In almost all conditions there is a huge disproportion between "Yes" and "No" answers. This allows to interpret them in a simple manner. *Before* seems to be treated as not EMON↑, as EMON↓, as PMON↑, not as PMON↓, and as ENTAIL↑. There is a problem with ENTAIL↓. When we feed this data into the format similar to the Table 3.2, we obtain such a line:

Definition	EMON↑	EMON↓	PMON↑	PMON↓	ENTAIL↑	ENTAIL↓
BEF_{exp}	No	Yes	Yes	No	**Yes**	??

EMON and PMON properties are in accordance with the definition BEF_0 but ENTAIL differs. I believe, that this (almost) validates this particular definition as a proper definition for a natural language *before*, but we need to explain why this difference is present.

There are two more things to notice before giving the explanation. First concerns the change in answers between the time lags. It was not as uniform in the case of *before* as it was in the case of *until*. I.e. for some conditions, there were more negative answers in the short lag, than in the long lag (e.g. EMON$\downarrow$ or ENTAIL$\uparrow$). Since these changes are very small, it doesn't seem to be anything more than noise.

The second one is a consequence of lack of differences between the time lags. This means, that either there weren't any pragmatic inferences triggered by *before*, or that, the ones present couldn't be eliminated by enlarging the processing load. The former is false. It can be easily observed, that *before* is a presupposition trigger in the temporal clause, and that the items used in measuring the ENTAIL properties carried a presupposition. The fact, that there weren't any differences between the time groups, might suggest that processing presupposition doesn't carry any additional cognitive cost. A tentative argument follows.

The results from EMON and PMON conditions show, that *before* does not behave in accordance with the definition BEF_1. Let's assume, that BEF_0 is the correct definition. Consider an ENTAIL$\downarrow$ item:

(1A) Tom from the middle group was drawing before going home.

(1B) Tom from the middle group was drawing before going home and playing football.

I believe that the sentence (B) presupposes, that there was a time in which Tom was playing football. Now, if processing presupposition would carry an additional cost, then at least some of the subject shouldn't process the presupposition in the short lag. For these subjects the (B) sentence would have different truth conditions in each time group. In the long lag these would be:

(i) there was a time t_1 in which Tom was drawing,

(ii) there was a time t_2 in which Tom went home and in which Tom was playing football,

(iii) for all such times t_2, $t_1 < t_2$.

In the short lag (ii) wouldn't hold. Knowing, that something is the case for all possibilities, but not knowing whether there exists a single instance of such possibility, is very different from knowing that something is the case for all possibilities, and there is an actual instance of such a possibility. This should have an impact on answers. The fact that it does not, indicates that the subjects do process presupposition in both time lags.

The argument above might not be very convincing, since there might be many other reasons why no difference is noted. But it goes along the way, which I think may explain the unexpected answering patterns in the ENTAIL conditions. It has to do with tacit presuppositions that are put forward in the experiment. *Before* is indeed a downward entailing operator which is best seen in the EMON conditions. Within those, the subjects were almost uniformly accepting the DE inference. In a sense, it would have been strange if they weren't. Consider another item:

(2A) Alex from the small group was laughing before fun and games.

(2B) Alex from the small group was laughing before games in groups.

The timetable explicitly states, that "games in groups" were a part of fun and games. Therefore, there existed such an event as "games in groups" and it was later than fun and games. Since the subject is informed about this, the only way of not accepting the inference is, by interpreting *before* as intransitive - which, as we can see, is rarely the case.

Because of the particular experimental setting, subjects are being accustomed to the fact, that all information have to be somehow confirmed by the timetable. With ENTAIL items this wasn't the case (compare 1A, B). However, in ENTAIL↑ the subjects were first reading a statement asserting an existence of two given events in the temporal clause. This made accepting the B sentence almost automatic - there was no reason to doubt the connection between the two events, which was needed to reject the inference.

In ENTAIL↓ the situation was different. The subject had to validate the existence of the presupposed event by herself (or himself). Since nothing in the setting told her/him that there was a second event, the subjects were hesitant to accept the consequence. The fact, that "Yes" and "No" answers are distributed almost equally shows, that even in a setting in which the subjects were discouraged to assume the existence of a given event, they were reluctant to reject the reasoning. Probably in a different setting the ENTAIL↓ property would be much more visible.

The analysis of entailment given above implicitly uses the notion of so called *Strawson Entailment* (Von Fintel [1999]). This is based on the notion of *Strawson-Validity*:

SE 1 (*Strawson-Validity*)**.** An inference $p_1, p_2, p_3, \ldots p_n \models q$ is *Strawson-Valid* iff the inference $p_1, p_2, p_3, \ldots p_n, S \models q$ is (classically) valid, where S is a premise stating that the presuppositions of all the statements involved are satisfied.

Von Fintel derives this notion from Strawson's proposal on how to make a classical inference, from *Every S is P* to *Some S is P* valid, within the framework of modern logic. As Von Fintel writes, Strawson's idea was to assume, that every natural language quantifier carried a (semantic) presupposition of existence with respect to their domain. So a *Strawson Downward Entailment* could be defined as:

SE 2 (*Strawson Downward Entailment*)**.** A function $\mathcal{F}$ is *Strawson-DE* iff for all p, q such that $p \Rightarrow q$ and $\mathcal{F}(p)$ is defined: $\mathcal{F}(q) \Rightarrow \mathcal{F}(p)$.

Given the experimental setting which drawn particular attention to the presuppositions, and the definitions above, I believe that assuming, that *before* in ENTAIL conditions is *Strawson-DE* explains the data obtained in these conditions.

Until

The results for *until* are much more interesting. This is mainly because of the observed differences between the time lags. We shall start by reflecting on the summarised results. *Until* seems to be not EMON↑, not EMON↓, not PMON↑, not PMON↓, probably not ENTAIL↓, but it is ENTAIL↑. Comparing these results with the Table 3.2 we obtain:

Definition	EMON↑	EMON↓	PMON↑	PMON↓	ENTAIL↑	ENTAIL↓
UNT_{exp}	No	No	No	No	Yes	No

These are exactly the conditions satisfied by the definition UNT_2. Note that this is not the widely assumed definition of *until*, but rather the one with the end point encoded in the semantics.

This very clean image is immediately shattered at the moment in which we move our attention to the results in the different time lags. The differences in PMON↑ and EMON↓ are clearly not accidental (see Table 3.6 for a remainder). The differences are in favour of "Yes" answer in the short lag. Such a result can be interpreted as indicating that "No" answer was more costly. As written before, it was repeatedly shown, that such a cost is connected to the pragmatic inferences such as scalar implicature (compare e.g. De Neys and Schaeken, Bott and Noveck). Given all this, we should interpret differences between the lags in PMON↑ and EMON↓ conditions as indicating that some sort of a pragmatic inference is playing a part.

Pragmatic meaning of an expression is triggered by its use and is not a part of its semantics. Therefore we should investigate what the proper semantics for *until* could be, assuming that the results in question are a consequence of pragmatic phenomena of some sort.

Unfortunately, simply treating the results, as if the proper definition for *until* would be EMON↓ and PMON↑ will not work, because there is no definition with such properties. Also note, that PMON properties in the definitions UNT_1 and UNT_2 follow from the constraints put on the temporal order of event arguments. If we do not want to alter the definition of *until* completely, we have to stick to either non-PMON or PMON↓-PMON↑ is of course neither. Similar remarks can be made about EMON properties.

Therefore, we should begin the analysis by examining the items from the crucial conditions (a relevant timetable is the one presented in Section 3.2.1 under "Variables and Materials").

PMON↑

(3A) Alex from the small group was jumping until games in groups.

(3B) Alex from the small group was jumping until lunch.

EMON↓

(4A) Sophie from the small group was telling stories until fun and games.

(4B) Sophie from the small group was telling stories until games in groups.

Obviously there is a similarity between those two items. In both cases, sentence A in the temporal clause refers to an event which precedes the relevant event from the sentence B. There is also another similarity: in both items the events mentioned in the temporal clause are co-temporal. In the case of PMON↑ the event "games in groups" ends exactly when the next event begins. In the case of EMON↓ the second event is a part of the first event.

This having said, one obvious interpretation appears. *Until* does not have an end point encoded in its semantics. However, such an end point almost always appears as an implicature which *until* triggers. So, when a subject encountered one of the aforementioned items in the long lag, she interpreted *until* as *until and no longer*. The differences in the short lag show, that sometimes the derivation of the upper bound was too difficult, and subjects were forced into literal reading, in which there is no end point, so the subjects couldn't decide, until what time exactly the event in question took place.

It is worth noting, that such a treatment of *until* is in line with the quite common interpretation of numeral modifier *up to*, which bears close resemblance to *until* (see e.g. Nouwen [2008], Blok [2015]). All this seems a strong argument in favour of the definition UNT_1.

However, there is one problem if we chose UNT_1 as the proper definition of *until*. This definition is EMON↑, but the subjects did not treat *until* as such. Why is that? Again, consider an EMON↑ item:[6]

(5A) Jacob from the small group was drawing until games in groups.

(5B) Jacob from the small group was drawing until fun and games.

Since 5A is true, we know that it means (see Intuitions 1.4):

(a) there is a time t_1 at which Jacob is drawing,

(b) there is a time t_3, at which games in groups takes place,

(c) t_3 succeeds t_1,

(d) for all such times t_2 which are in between t_1 and t_3, Jacob is also drawing in t_2

This might seem odd at first, but none of the conditions (a) - (d) ensures that 5B must be true when 5A is. Recall that *fun and games* are earlier than *games in groups*. We only know that Jacob started drawing at some point which was earlier than *games in groups*, but this does not necessarily means that it was earlier than *fun and games*. He might have started drawing during *fun and games*. This is particularly strange, since we've shown that UNT_1 *is* EMON↑. It seems that our logic can not express such counterexample.[7]

Supposedly the same applies to PMON↓. Such a result is quite interesting, since it shows that the logical analysis presented above is too coarse grained to account for all the features of the discussed operators. We shall now proceed to the last part of the discussion.

[6]This explanation was suggested to me by Rick Nouwen during the presentation at the ROSE seminar.
[7]For instance, take the structure described under *A working example* in Section 2.2. Note that $f \sqsubseteq e$ is similar to our case of *games in groups* being a subevent of *fun and games*. But neither $UNT_1(e, f) \rightarrow UNT_1(e, e)$ nor $UNT_1(f, f) \rightarrow UNT_1(f, e)$ aren't counterexamples.

3.3.2 Possible confounds

Most of the confounds were discussed above as a part of the interpretation. The most serious ones concern the particular design of the experiment.

First, the way the schedule was presented (via the timetables), inflicted a very specific temporal ontology on the subjects. The events were of a discrete nature, and they had a clear start and end points. One might argue, that time is dense and temporal prepositions we use, somehow reflect this. Since the setting in the experiment clearly isn't dense, the use of the temporal operators might be distorted. However, even if this had any real influence on the results, it didn't affect what we wanted to measure. Inferential properties shouldn't be susceptible to any of the factors mentioned above.

As was already mentioned, suspiciously big amount of subjects answered the control questions incorrectly. This could be because of various reasons. First of them concerns the names used. Throughout the paper only several names (with group indicator) were used, so the subjects could put up a mental representation of each child. This was meant to help answer the questions, but it could have unwanted influence which might show in answers to control questions. E.g. if one gets particular information about the Suzy from the middle group, they might feel they now better whether she had dinner at one o'clock.

The second reason might be connected to the tense of the sentences used in the items. Sentences in the target items were put in past tense, in the control items this was present tense. Possibly this could influence subjects' readings by suggesting that these facts differ in some aspect from the ones in past tense. Such contrast may provoke hesitation: what happened in the past was given as a fact, what happens in the present needed interpretation. This noise could have influenced the results, but it is difficult to judge how much.

Another objection is rather quite a general one. Possibly no linguistic item exists without a reason. In Polish, there are at least three ways of expressing the relations denoted by *before* and *until*. Usually the use of these expressions is not as uniform as it was in the experiment. Because of this, some of the items might not have been as natural, as they would be in spoken language. This does not seem to be a very serious objection. The results show, that we managed to get reasonable answers.

The last remark probably couldn't have any influence on the results, but is worth stating for some other reasons. Judging on the feedback from the participants, the experiment was long and tiresome. One might argue, that experiments exploring human cognitive abilities, have to be cognitively demanding. I believe, that for the sake of future research, one should make them as easy and pleasant as possible.

Chapter 4

Conclusion

In the paper several things were obtained. First, I described the aim of the work and stated motivations. Then, a logical event-framework (due to Kamp [1979]) was presented. In this framework, I reconstructed some definitions for two temporal operators (*before* and *until*) in a uniform way. Three types of monotonicity properties for these operators were proposed. For each of the operators, certain monotonicity properties were shown. In turn, this allowed for stating precise experimental predictions.

I tested these predictions in a study using a paradigm similar to the one devised by Chemla and Bott [2014]. In addition, the subjects' use of cognitive abilities was controlled by introducing two time groups. The results showed differences between the two operators.

Before seems to be a downward entailing operator which triggers presuppositions - this is best explained by using the notion of *Strawson Entailment*. *Until* seems to be a more complex operator, which has a lexicalised pragmatic component. The differences within the operators were revealed, by varying the cognitive load the subjects had to endure. This shows, that even when using simple temporal prepositions, many different cognitive processes are involved.

Appendix A

Period Structures

Another way of obtaining instants is through an intermediate *period* structure. Define equivalence relation on events in such a way, that one event is equivalent to the other when they go on at the same time. Formally:

P 1. $\approx = \lambda e \lambda e'.e \sqsubseteq e^{'} \wedge e^{'} \sqsubseteq e$

So a period will be an equivalence class over events. A set of periods is a set of equivalence classes under this relation: $[E]_{\approx}$. Because of this, instead of writing (e.g.): $\pi_1 = \{a\}$ we could also write: $[a]_{\approx}$. Since periods are equivalence classes over events, there is quite a natural way of creating a structure over such periods.

P 2. $\pi_1 < \pi_2$ iff $\exists e \in \pi_1 \exists e' \in \pi_2 (e < e')$

One period precedes the other when an underlying event of the first period precedes an event in the other period.

P 3. $\pi_1 \bigcirc \pi_2$ iff $\exists e \in \pi_1 \exists e' \in \pi_2 (e \bigcirc e')$

Similarly, one period overlaps the other when an underlying event of the first period overlaps an event in the other period.

P 4. $\pi_1 \sqsubseteq \pi_2$ iff $\exists e \in \pi_1 \exists e' \in \pi_2 (e \sqsubseteq e')$

And finally, one period is a temporal subset of the other when an underlying event from the first period is a temporal subset of an event in the other period.

This way we have obtained a *generated period structure*:

$$\mathbb{P} = \langle [E]_{\approx}, <, \bigcirc, \sqsubseteq, \rangle$$

.

From such structure we can get instants in a similar way to what was shown above. We shall define *quasi-filters* which corresponds to instants.

A *quasi-filter* in $\mathbb{E}$ is a set $F \subseteq \mathbb{E}$ such that: $\forall \pi_1 \forall \pi_2 \in F(\pi_1 \bigcirc \pi_2)$.

We shall constrain ourselves only to quasi-filters which are non-empty and which can not be extended any further to a quasi-filter. These are called *maximally consistent quasi-filters*. Let:

$$\mathbb{I} = \langle I, < \rangle \text{ where:}$$

- I is a set of all maximally consistent quasi-filters in $\mathbb{P}$
- $i_1 < i_2$ iff $\exists \pi_1 \in i_1 \exists \pi_2 \in i_2(\pi_1 < \pi_2)$

We could convince ourselves that this structure is quite the same (isomorphic?) as the previous one by considering the enriched working example from the section 2:

Let $E = \{a, b, c, d, e, f, g, h\}$. Let relations between events in E be:

- Temporal precedence: $a < c, c < e, d < f, f < g$
- Subevent relation: $a \sqsubseteq b, b \sqsubseteq a, g \sqsubseteq h, h \sqsubseteq g$
- Overlap relation: $c \bigcirc d, d \bigcirc e, e \bigcirc f$

Periods: $\pi_1 = \{a, b\}, \pi_2 = \{c\}, \pi_3 = \{d\}, \pi_4 = \{e\}, \pi_5 = \{f\}, \pi_6 = \{g, h\}$.
Time points: $t_1 = \{\pi_1\}, t_2 = \{\pi_2, \pi_3\}, t_3 = \{\pi_3, \pi_4\}, t_4 = \{\pi_4, \pi_5\}, t_5 = \{\pi_6\}$.
Again: $t_1 < t_2 < t_3 < t_4 < t_5$.

Note, that when using this approach the definitions for *before* and *until* have to be altered in such a way which would take periods into account. E.g.:

$$BEF_0'(a, b) \equiv_{\text{df}} \exists t_1[([a]_{\approx} \in t_1) \wedge \forall t_2(([b]_{\approx} \in t_2) \rightarrow (t_1 < t_2))]$$

Appendix B

Experimental Items

Below the way of obtaining experimental items is presented. Recall that the experiment was conducted in Polish, below the reader may see a translated version.

Target Items and Fillers

All target items were of the form:

$$X \ [\textit{activity}] \ [\textit{temporal operator}] \ [\textit{timetable item}]$$

Where:

- X was the name of a child with the indication to which group it belong (there were two groups: *the middles* and *the smalls*). The following names were used: Jacob, Sophie, Alex, Mary, Suzy, John, Tom, Peter. First four were used in the small group, the rest in the middle group.
- Examples of activities used: drinking water, playing tag, laughing, drawing, running, playing with dolls, coughing.
- Temporal operators were either *before*, or *until*.
- For a timetable used in the small group see Section 3.2.1. The other one was constructed in a similar manner.

Fillers were constructed in a similar way, but different temporal prepositions were used (*after*, *while*).

Controls

Sentences used in control items differed accordingly to whether they were used as an assumption or as a consequence. The sentences used as assumption were all of the form:

It is [*hour*]. [*name*] is in the [*group*]

The sentences used as consequence were all of the form:

[*name*] [*group indicator*] is [*doing*] [*timetable item*]

name, group indicator and *group* were the same for each control item. *doing* indicates a verb used to describe that the child in question is taking part in particular activity from the timetable. Controls were balanced for truth and falsity.

Bibliography

William Allen Ladusaw. *Polarity Sensitivity as Inherent Scope Relations*. PhD thesis, University of Texas, 1980.

Gilles Fauconnier. Polarity and the scale principle. In *Papers from the Eleventh Regional Meeting of the Chicago Linguistic Society*, pages 188–199. Chicago Linguistic Society, Chicago, IL, 1975.

Jon Barwise and Robin Cooper. Generalized quantifiers and natural language. *Linguistics and philosophy*, 4(2):159–219, 1981.

Michael Israel. The pragmatics of polarity. In Laurence Horn and Gergory Ward, editors, *The Handbook of Pragmatics*, pages 701–723. Blackwell, Oxford, 2004.

Ido Dagan and Oren Glickman. Probabilistic textual entailment: Generic applied modeling of language variability. In *PASCAL Workshop on Learning Methods for Text Understanding and Mining*, pages 26–29, Grenoble, France, 2004.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. *Recognizing textual entailment: Models and applications*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2013.

Luisa Bentivogli, Peter Clark, Ido Dagan, Hoa Dang, and Danilo Giampiccolo. The seventh pascal recognizing textual entailment challenge. In *Proceedings of the Fourth Text Analysis Conference (TAC 2011)*, 2011.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. Don't'have a clue'?: unsupervised colearning of downward-entailing operators. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 247–252. Association for Computational Linguistics, 2010.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Richard Ducott. Without a "doubt"?: Unsupervised discovery of downward-entailing operators. In *Proceedings of Human*

Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 137–145. Association for Computational Linguistics, 2009.

Hans Kamp. Events, instants and temporal reference. In *Semantics from different points of view*, pages 376–418. Springer, 1979.

Bertrand Russell. *Our knowledge of the External World as a Field for Scientific Method in Philosophy.* Open Court, 1914.

Friedrich Hamm and Oliver Bott. Tense and aspect. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Stanford University, spring 2014 edition, 2014.

Jeffrey M Zacks, Barbara Tversky, and Gowri Iyer. Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General*, 130(1): 29–58, 2001.

Joost Zwarts and Yoad Winter. Vector space semantics: A model-theoretic analysis of locative prepositions. *Journal of logic, language and information*, 9(2):169–211, 2000.

Gertrude Elizabeth Margaret Anscombe. Before and after. *The Philosophical Review*, pages 3–24, 1964.

David Beaver and Cleo Condoravdi. A uniform analysis of "before" and "after". In *Proceedings of the 13th Semantics and Linguistic Theory Conference*, pages 37–54, 2003.

Fred Landman. *Structures for Semantics.* Kluwer Academic Publishers, 1991.

Orvokki Heinämäki. Before. In *Papers from the 8th Regional Meeting of the Chicago Linguistic Society*, pages 139–51. Chicago Linguistic Society, Chicago, IL, 1972.

Johan Anthony Wilem Kamp. *Tense logic and the theory of linear order.* PhD thesis, University of California, Los Angeles, 1968.

Rick Nouwen. Directionality in numeral quantifiers: the case of "up to". In *Proceedings of the 18th Semantics and Linguistic Theory conference*, pages 569–582, 2008.

Johan Van Benthem. Questions about quantifiers. *The Journal of Symbolic Logic*, 49 (02):443–466, 1984.

Henriette De Swart. Meaning and use of not ... until. *Journal of Semantics*, 13(3): 221–263, 1996.

Ira A Noveck. When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, 78(2):165–188, 2001.

Lewis Bott and Ira A Noveck. Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of memory and language*, 51(3):437–457, 2004.

Yi Ting Huang and Jesse Snedeker. Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive psychology*, 58(3):376–415, 2009.

Emmanuel Chemla. An epistemic step for anti-presuppositions. *Journal of Semantics*, 25(2):141–173, 2008.

Coralie Chevallier, Ira A Noveck, Tatjana Nazir, Lewis Bott, Valentina Lanzetti, and Dan Sperber. Making disjunctions exclusive. *The quarterly journal of experimental psychology*, 61(11):1741–1760, 2008.

Herbert Paul Grice. *Studies in the Way of Words*. Harvard University Press, 1989.

John Langshaw Austin. *How to Do Things with Words*. Oxford University Press, 1962.

Ludwig Wittgenstein. *Philosophical Investigations*. Blackwell, Oxford, 1953.

Bart Geurts. Reasoning with quantifiers. *Cognition*, 86(3):223–251, 2003.

Bart Geurts and Frans van Der Slik. Monotonicity and processing load. *Journal of semantics*, 22(1):97–117, 2005.

Carrie O'Leary and Stephen Crain. Negative polarity items (a positive result) positive polarity items (a negative result). Paper presented at the 1994 Boston University Conference on Language Development, 1994.

Julien Musolino, Stephen Crain, and Rosalind Thornton. Navigating negative quantificational space. *Linguistics*, 38(1):1–32, 2000.

Makoto Kanazawa. Weak vs. strong readings of donkey sentences and monotonicity inference in a dynamic setting. *Linguistics and philosophy*, 17(2):109–158, 1994.

Bart Geurts. Donkey business. *Linguistics and Philosophy*, 25:129–156, 2002.

Laurence Horn. *A Natural History of Negation*. CSLI Publications, Stanford, 1989.

Peter C. Wason. Reasoning about a rule. *The Quarterly journal of experimental psychology*, 20(3):273–281, 1968.

Emmanuel Chemla and Lewis Bott. Processing inferences at the semantics/pragmatics frontier: disjunctions and free choice. *Cognition*, 130(3):380–396, 2014.

Stephen C. Levinson. *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press, 2000.

Dan Sperber and Deirdre Wilson. *Relevance: Communication and Cognition*. Harvard University Press, 1986.

Wim De Neys and Walter Schaeken. When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental psychology*, 54(2):128–133, 2007.

Richard Breheny, Napoleon Katsos, and John Williams. Are generalised scalar implicatures generated by default? an on-line investigation into the role of context in generating pragmatic inferenceson-line investigation into the role of context in generating pragmatic inferences. *Cognition*, 100(3):434–463, 2006.

Kai Von Fintel. Npi licensing, strawson entailment, and context dependency. *Journal of semantics*, 16(2):97–148, 1999.

Dominique Blok. The semantics and pragmatics of directional numeral modifiers. Unpublished Manuscript, 2015.