Unsaid Thoughts: Thinking in the absence of (some) verbal logical connectives

David J. Lobina,[1*] Josep Demestre,[2] José E. García-Albea[3] & Marc Guasch[2]

[1]Psycholinguistics Research Group, Universitat Rovira i Virgili
[2]Department of Psychology and Research Center for Behavior Assessment,
Universitat Rovira i Virgili, Ctra. de Valls s/n, 43007 Tarragona, Spain
[3]Department of Psychology, Universitat Rovira i Virgili,
Ctra. de Valls s/n, 43007 Tarragona, Spain

[*]To whom correspondence should be addressed; E-mail: djlobina@gmail.com

Combining two thoughts into a compound mental representation is a central feature of our verbal and non-verbal logical abilities. We here approach this issue by focusing on the contingency that while natural languages typically verbalise only two of the sixteen connectives from formal logic to express compound thoughts —*and* and *or*— the remainder appear to be entertainable as non-verbal, conceptual representations and this suggests a way to probe how linguistic and non-linguistic thinking processes relate. In a *visual world* experiment aimed at tracking both comprehension-related and reasoning-related aspects of the capacity to represent compound thoughts, we found that participants are capable of learning and interpreting a made-up word for logic's NAND operator, indicating that unlexicalised logical connectives are nonetheless conceptually available.

A core feature of cognition is the ability to combine two propositions (or thoughts) into com-

plex mental representations (*1, 2*), and a principal way to examine this capacity is by employing the tools of formal logic, a field of study that has informed our understanding of human reasoning since antiquity and which remains prominent in contemporary cognitive science (*3, 4*). The input from logic has been especially fruitful in experimental work on how we represent, and reason with, compound propositions, including the complementary ability to generate further thoughts from such propositions (i.e., to draw conclusions from complex propositions) (*5, 6*).

Much of the evidence has come from studies using linguistic representations of propositions, given that complex propositions can be often expressed linguistically through so-called coordinators (*7*), as in the compound sentence *the triangle is yellow AND/OR the square is blue*, where each clause (*the triangle is yellow*, *the square is blue*) constitutes a proposition and the coordinators "and/or" function as the sentential operators, or connectives, from logic. This kind of approach has necessarily often strayed into the rather thorny question of how the human capacity for language relates to the ability to think and reason, and as we shall argue here, the study of language's logical connectives can indeed highlight both verbal and non-verbal reasoning processes (and their interaction).

The use of linguistic representations as a proxy to probe how we reason with compound propositions can give rise to some complications, however. A notable problem stems from the nature of language itself and of linguistic communication in particular: compound sentences are underlain by properties of various kinds and these can all play a role in the interpretation and use of such sentences in actual exchanges. Simplifying somewhat, these features can be split into two types, the linguistic and the non-linguistic. Relevant linguistic properties include the syntax of a sentence as well as the semantic and pragmatic information sentences codify, while on the side of non-linguistic properties we especially note that compound sentences can give rise to *rules of inference*, logical rules that are not explicitly encoded in the sentences themselves but which would have an effect on how such sentences are comprehended and used.

Unsurprisingly, and given the way in which these properties can interrelate, an experimental investigation evaluating how compound sentences are processed and indeed logically interpreted needs to pay especial attention to these features.

On the linguistic side of things, compound sentences very often appear to behave in ways that diverge from what is the case in logic (*8*); conjunction *and*, for instance, can signal much more than a simple union of propositions, which is what logic mandates (e.g., it can mark a temporal, and even causal, relationship between two clauses, as in *the bomb exploded and the car was destroyed*). The consensus in the literature, however, is that the meaning —the semantics— of some compound sentences (and thus of some linguistic coordinators) is analogous to the meanings logicians assign to compound propositions and connectives in terms of truth tables (see Table 1), with the corollary that the non-logical uses the relevant complex sentences exhibit in language use are the result of diverse pragmatic processes (presuppositions, implicatures, etc.). To be more precise, the consensus is based on the contingency that for the most part pragmatic effects can be nullified in various ways, thereby unearthing the logical meaning of the applicable compound sentences. Among other ways, pragmatic effects can be cancelled by the context or the addition of further linguistic material (*9*), they can be blocked by specific syntactic configurations (*10*), and some of them can even be controlled for in an experimental setting (*11*), the latter the approach we used in this study (see *Supplementary Materials*).

Thus, the meaning of language's *and* corresponds to the truth table of logic's conjunction ($\wedge$, shown in column 3 of Table 1) —viz., a conjunctive compound sentence is only true when both clauses are true. A similar situation applies to language's *or*, though the state of affairs in this case is more intricate. Even though language's *or* often receives an "either/or" reading in linguistic exchanges, which would entail that its meaning corresponds to logic's exclusive disjunction ($\vee_e$; column 6) —i.e., a disjunctive sentence would be true only if either one of the

two clauses is true— its actual, default meaning, as ascertained both empirically and through various linguistic tests, is in fact intrinsically inclusive (*12–14*). That is, language's *or* specifies an "P or Q, or both" interpretation, and therefore its (semantic) meaning more properly matches the truth table of logic's inclusive disjunction ($\vee$; column 5) —a disjunctive compound sentence is true if one of the two clauses is true as well as when both clauses are true (it is important to note that the truth table of exclusive disjunction constitutes a subset of the truth table of inclusive disjunction, both in logic and in language). The either/or reading seemingly so common in linguistic exchanges is argued to be the result of an "exclusivity implicature" hearers tend to compute in real-time communication.

Table 1: Logic's Truth Tables. Columns 1–2 show propositions P and Q and the different combinations of the truth values logicians assign to them in terms of whether the propositions describe a state of the world or not (T for true, F for false). The remaining columns specify the truth tables for the logical connectives Conjunction and NAND (columns 3–4), Inclusive and Exclusive Disjunctions (5–6), and NOR (7).

P	Q	$\wedge$	NAND	$\vee$	$\vee_e$	NOR
T	T	T	F	T	F	F
T	F	F	T	T	T	F
F	T	F	T	T	T	F
F	F	F	T	F	F	T

On the side of non-linguistic, reasoning-related properties of compound sentences, a disjunctive sentence such as *the triangle is yellow or the square is blue* can sometimes raise the question of whether only one of the clauses is in fact true, thus requiring that the other clause be discarded (*15*). This interpretation usually surfaces when a disjunctive sentence is presented as the premise to a problem-solving task; in such circumstances, participants appear to apply a "disjunction elimination" strategy. This rule of inference determines, that when presented with premises *P or Q* and *Not P*, it follows that *Q* applies (or is true); it has been investigated in both verbal and preverbal children (*16, 17*) and both adults and very young children appear to conform to it.

4

Inference rules of this sort are plausibly the result of post-linguistic processes, given that they tend to materialise when participants are asked to reason with compound sentences (such rules are not automatic or mandatory in any way). From the perspective of language processing, in fact, inference rules need not apply at all; a disjunctive sentence may potentially describe three different states of affairs, in accordance to its truth table, but in regular linguistic exchanges the comprehension system regularly computes a single interpretation only and with no need for any reasoning strategy to do so. The overall context as well as the nature of linguistic exchanges help hearers establish the intended interpretation, quickly so and with little, if any, reflection.

The relationship between linguistic and non-linguistic properties of compound sentences is necessarily a subtle affair, not least because in order to reason with these sentences adequately they must be comprehended appropriately. That is, the language comprehension system must be able to access all possible readings of compound sentences when required to do so if whatever mental systems are in charge of reasoning are to successfully apply compound sentences as premises to an inference rule. In such cases, the mapping between linguistic and the relevant reasoning systems can be rather blurry, though the respective processes certainly differ in important ways. Language comprehension is mostly an implicit process not available to introspection, while the ability to represent all possible interpretations of compound sentences and indeed reason with such readings is more explicit in nature. After all, it is clearly possible to reflexively consider what such sentences mean as well as what may follow from them, even if the actual details of how these interpretations and consequences are obtained are not directly accessible.

The overall picture illustrates the fact that there are different facets to the interface between the processes involved in the comprehension of language's logical connectives and the processes at play in the representation of all possible readings of compound sentences, and these aspects need to be carefully considered in an experimental investigation of the language-and-thought

interface. With this purpose in mind, we put together a tailor-made experiment to probe some of the implicit processes underlying the (fast) comprehension of compound sentences as well as some of the more explicit (and slower) phenomenon of reflecting upon the different interpretations compound sentences yield. In particular, we combined the visual-world paradigm with a sentence/picture matching task, employed a seldom used but potentially fruitful statistical model to analyse eye-tracking data, and exploited the peculiar way in which the logical connectives are realised and used in language.

The last point is central to the framework we implemented. It is noteworthy that even though there are a possible sixteen binary connectives in logic and language makes use of a great number of coordinators to link up different clauses (*and*, *because*, *if...then*, etc.), only two coordinators behave like logic's connectives. That is, of the sixteen possible logical connectives, only two have been unambiguously lexicalised in the world's languages: the aforementioned conjunction *and* and (inclusive) disjunction *or* (*18*), resulting in the relevant types of logical compound sentences, conjunctive and disjunctive sentences (we put to one side the issue of cross-linguistic variability in order to simplify matters, but in any case our discussion is uncontroversially true of the language of the experiments (*11*)).

This is not to say that natural languages lack the resources to express or describe the possible state of affairs that the sixteen logical connectives can account for; indeed, the truth tables of non-verbalised connectives can be derived analytically through the combination of *and*, *or*, and negation, a fact of language as much as of formal logic (*19*), though this would often produce rather convoluted sentences. What's more, all sixteen truth tables can in fact be derived by the repeated application of one single operator from logic, either the alternative denial connective, also known as NAND or Sheffer's stroke (column 4 in Table 1), or the joint denial connective, often referred to as logical NOR (column 7). In either case, the representation of certain meanings would require significantly long strings of derivations, resulting in a rather inefficient

6

medium of communication, which may partly explain why these two connectives have not been lexicalised in any language (the English word *nor* should not be confused with logical NOR, though there might be some relationship between the two (*20*)).

There has been plenty of discussion in the literature as to why natural languages have verbalised a specific set of connectives and not another; or said otherwise, why some connectives have been blocked from being lexicalised. Most explanations have pointed to so-called economy considerations regarding the derivability of the non-verbalised connectives, in addition to a restriction on "negative" connectives —NAND and NOR are examples of such connectives, as the former is the contrary, or opposite, of conjunction *and* and the latter of inclusive disjunction *or*, as can be ascertained in Table 1 by contrasting columns 3 and 4, on the one hand, and columns 5 and 7, on the other (the truth tables are reversed, as it were) (*12, 18, 21*). Crucially, these "blocking effects" do not actually establish that the unlexicalised connectives constitute so-called impossible words (*22*); they simply attempt to explain why some connectives and not others have been lexicalised. Some concepts do appear to be genuinely impossible to lexicalise —i.e., to become established words of a language, something that is especially true of some verbs— perhaps for metaphysical reasons (*23*) or on account of intrinsic linguistic constraints (*22*), but the concepts themselves seem perfectly entertainable in a conceptual representational system of the mind, a *language of thought* (*2*). This appears to be the case for the logical connectives as well, the very factor we focused on in this study.

This particular question has not received as much attention in lexicalisation accounts of the connectives, and yet it suggests a potentially valuable way to explore non-verbal properties of cognition. The key point is that a lexicalisation account does not preclude the possibility that unlexicalised connectives may be learned; that is, if there are no reasons to believe that unlexicalised connectives constitute impossible words, then it ought to be possible to devise an experimental task in which participants would be expected to learn and appropriately compre-

hend made-up words standing for unverbalised connectives. In other words, if a concept is entertainable and it *could* have been lexicalised had the facts of the matter been conducive to it, then a word meant to embody the meaning of such a concept could be learned in an appropriate experimental setting. A theory of the lexicalisation (or lack thereof) of the connectives is not *ipso facto* an account of the learnability (or not) of non-existent words, pointing to a theoretical split between lexicalisation and learnability (*24*).

The unlexicalised connectives NAND and NOR are good candidates for such an undertaking, given their connection to the existing linguistic connectives *and* and *or*, and the fact, derived from this tie, that they do not express especially convoluted state of affairs —e.g., neither is only true in case the first clause is true but false in any other instance, which would be an odd linguistic structure indeed (such a meaning corresponds to the truth table of the contrary of material implication, naturally an unlexicalised connective).

In this study we specifically focused on NAND, as its truth table specifies three possible situations, for just one for NOR, and thus the learning of this connective would be a clearer demonstration that the requisite unlexicalised concept is not only entertainable but that it can furthermore be appropriately applied to a non-existent word. Indeed, in comparing the difficulty of learning new words for NAND and NOR, the latter might be too easy a task, as a strategy that simply negates the two clauses of the presented compound sentences would suffice to show competence with the meaning of the made-up word for NOR. After all, in order to demonstrate that the NAND concept is available and moreover applicable to a novel word, participants would have to accept the truth of a NAND compound sentence when either one of the two clauses is true (and the other false), in addition to when both clauses are false, and this would require a more nuanced interpretation than in the case of NOR.

We employed the visual-world paradigm in combination with a behavioural sentence-picture matching task, as described in Figure 1, Panel A. The visual-world set-up was slightly different

from what is usually the case in the field; a typical design uses four regions or areas of interest on a computer screen, each displaying a graphic, in order to use one graphic as the target interpretation for some aspect of the sentence played to participants, one graphic as a potential competitor, and two further graphics as distractors. In our set-up, the graphic on each quadrant was instead a representation of a possible combination of the two values from each line of a truth table —namely, the combinations TT, TF, FT and FF (columns 1–2 from Table 1)— and thus which combinations would properly match the sentences presented to participants —i.e., which graphic would constitute a true statement of the graphics— would depend on which connective is used to put together the two clauses. In the case of conjunctive sentences, for instance, only the quadrant representing the TT combination would be the right interpretation for these sentences, while three such quadrants would match conjunction's opposite, the NAND connective (as mentioned, quadrants TF, FT, and FF).

We carried out two experiments; in addition to a task with NAND sentences, the target of the study, an experiment with disjunctive sentences was also conducted, and three quadrants were applicable in this case too (namely, TT, TF, and FT). The overall study was thus divided into two main experiments, which were administered in the same session and in order; in the first part of the session participants would do the task described in Figure 1, Panel A with disjunctive sentences and in the second part the task would be carried out with compound sentences in which the two clauses would be mediated by a non-existent but possible word standing for NAND. The disjunction experiment was included to familiarise participants with the task so that there were no extraneous effects of any kind during the NAND experiment; disjunctive sentences also constituted a good control for both the analyses we intended to conduct and participants' performance with NAND sentences (see the **Supplementary Materials** for details). The NAND session was itself divided into two parts. Participants would start this session by undertaking a learning phase in which they would be shown a series of situations that NAND sentences

9

appropriately describe (or not, as in the case of the TT reading, the one *false* interpretation for NAND sentences). Once the learning phase was completed, participants would carry out the same kind of task they had performed in the experiment with *or* (see **Materials and Methods** for more details).

As described in Figure 1, Panel A, the underlying idea of the study was to track participants' eye movements as the compound sentences are aurally presented, from the start of the audio file until a few seconds after the audio has finished, at which point participants were asked to select as many quadrants as they thought properly matched the sentence they had just heard. Such an experimental procedure would produce three relevant pieces of information. The eye-tracking data during the sentence is played, including when the sentence ends, would provide an online, implicit record of language comprehension processes. The participants' eye movements immediately after the sentence has finished and for a period of a few more seconds afterwards would plausibly coincide with the more explicit processes related to working out the various interpretations the compound sentences allow. And finally, the forced choice responses at the end of the task would reflect the actual reflexive process of deciding which readings are definitely warranted. Put together, these different data would provide a fuller picture of the comprehension and reasoning processes involved in entertaining and linguistically expressing complex mental propositions.

Regarding the eye-movement data, participants were expected to eventually zero in on what might well be regarded as the default meanings of disjunctive and NAND sentences, notwithstanding the fluctuations that were likely to occur during each time-series, as the truth tables of both disjunction and NAND contain three true combinations. Thus, in the case of disjunction we expected participants to converge to the TT reading, following data from previous studies (*11*), even if overall attention was anticipated to be divided between the TT reading and the mixed forms TF and FT during and immediately after the sentences were played, as per the

10

expectation that the exclusive reading of disjunction is quite prominent in communication. As for NAND, we hypothesised that the preferred reading would be the FF interpretation, possibly the core meaning of a connective that is effectively the negation, or contrary, of *and*, though the mixed forms TF and FT would need to be considered too if NAND sentences are fully understood and indeed interpreted logically. Regarding the behavioural responses, we predicted that if participants were to demonstrate mastery of the full meaning of disjunctive and NAND sentences, the most common pattern of responses would be TT-TF-FT for disjunction *or* and TF-FT-FF for NAND (that is, patterns TTTF and FTTT, respectively), in accordance to their truth tables.

The overall set-up presented some analytical challenges in the case of the eye-tracking data, however. The quadrants participants would be exposed to, showcasing the TT, TF, FT, and FF truth combinations, were all potentially relevant and thus fixations on all four areas of interest needed to be tracked —and for a fairly long period of time in each trial to boot (roughly, 6 seconds), as we were particularly interested in unearthing when exactly during the time course of a trial do differences between quadrants in fact surface. Such a long record of eye-movement data was also likely to give rise to a high amount of autocorrelation in the data, a common occurrence in experiments dealing with time-series (*25*). And finally, an initial inspection of the data suggested a non-linear pattern, as indicated by the wiggliness of the lines representing proportions of fixations shown in Figure 1, Panel B, generated prior to the analyses. Thus, a commonly-used technique to analyse eye-tracking data such as the linear regression implemented in a *growth curve analysis* (*26*) was not adequate for the purposes of the study and other techniques used in the past are not methodologically sound (*27*).

We employed a *generalised additive mixed-model* (GAMM) (*28*) to account for these issues, a technique that is becoming more common in analyses of time-series data (*25, 29*), including a few eye-tracking studies (*30*). GAMMs are specifically useful on account of three features: the

models relax the assumption of a linear relationship between predicting variables and response variables by implementing so-called "smooth functions"; the autocorrelation of the data can be accounted for by the inclusion of an autoregressive model; and, particularly convenient for the aims of our study, the interpretation of these models is partly determined visually, allowing us to plot, among other things, the differences between conditions through time (see **Supplementary Text** for further discussion).

The overall data are shown in Figure 1, Panels B–C. According to the eye-movement record, participants eventually converged on the TT interpretation for *or*, though the mixed form TF also received a significant number of fixations, while for the NAND condition participants preferred the FF reading from very early on and did not divert from this interpretation. In the case of disjunction *or*, the best-fit GAMM included smooth functions for time per condition (TT, TF, FT, and FF), the fixed effect, as well as smooths of time per time-series (i.e., per trial) per condition (TT, TF, FT, and FF), accounting for random effects, and in every case non-linear curves were obtained, as confirmed by the effective degrees of freedom (edf), a summary statistic of GAMMs that reflects the degree of non-linearity of a curve. An edf equal to 1 corresponds to a linear relationship between the predicting variables and the response variable, and anything above 2 equals to a highly non-linear relationship, which is what was observed in this model for every one of the eight smooth functions (with a $p < .001$ in every case). Figure 2, Panel A shows the "difference curves" between the smooth function for TT, the most fixated quadrant, against the remaining three smooths (for TF, FT, FF). Difference curves are the appropriate way to evaluate differences among the various levels of an experimental condition in GAMMs, in this case the main fixed effect of "area of interest" (or quadrant), showcasing the importance of the visual inspection of the data in these models. Indeed, these curves allow researchers to plot when during trials the estimated differences actually arise, and Figure 2 both marks and describes the relevant time windows for each comparison in each experiment (note the relative

variability from Experiment 1, where the TT-TF comparison is the most important one).

In the case of NAND, the best-fit GAMM included smooth functions for time per condition, the fixed effects, as well as by-participants and by-items smooth functions for time per condition, the random effects, for a total of 12 curves. Here too the non-linearity of all curves was confirmed, as the edf value was above 2 in every case (again, with $ps < .001$). Regarding the difference curves between the smooth for FF, the most fixated quadrant, against the smooths for FT, TF, and TT, and as shown in Figure 2, Panel B, participants settled on the FF interpretation soon after the connective appeared and they kept their fixations on that quadrant for the remainder of the trial (see **Materials and Methods** for the full details of the analyses of both experiments, which we have simplified a great deal here).

Regarding behavioural responses, shown in Figure 1, Panel C, the preferred pattern of response was TTTF for disjunction *or* (that is, participants selected the TT, TF, and FT quadrants), and FTTT for NAND (participants selected TF, FT, and FF), in line with the respective truth tables. For the analysis of these data, we drew a distinction between acceptable (or correct) patterns of response and unacceptable (or incorrect) patterns in order to run chi-squared tests between the expected and observed responses, in two steps: first between acceptable and unacceptable responses as a way to confirm that the sentences had been interpreted correctly, and then within acceptable responses between the two acceptable patterns of interpretation we had identified for both disjunction and NAND.

In the case of disjunction *or*, we took the patterns TTTF and FTTF to be the only acceptable responses, as these are the truth tables of inclusive and exclusive disjunction, respectively, and everything else was regarded as a mistake in interpretation. The percentage of correct answers thus amounted to 67.41, for 32.59 of incorrect answers, and the difference between these two frequencies was significant ($\chi^2(1) = 27.16, p < .001$). Once it was ascertained that the task had been appropriately carried out by the participants —the proportion of correct answers was

greater than the proportion of incorrect answers— we compared the two patterns of acceptable responses. In this case, the percentage of responses for the inclusive interpretation of disjunction (TTTF) was 80.79, for 19.21 for the exclusive interpretation (FTTF), and this difference was clearly significant as well ($\chi^2(1) = 57.29, p < .001$).

The situation was slightly more nuanced for NAND. Whilst the full interpretation of this connective would correspond to the FTTT pattern, the FFFT response (that is, only quadrant FF is selected), properly speaking the truth table of the (unlexicalised) connective NOR, cannot be regarded as an entirely incorrect or unacceptable response for NAND, but perhaps simply an incomplete one. As noted, NAND is the opposite of conjunction, and in this sense the FF-only interpretation is plausibly the most straightforward counterpart to the TT reading of conjunction —or the default, preferred interpretation, as the eye-movement data in fact indicate, justifying the inclusion of FFFT as the second acceptable response along with FTTT. Thus established, the overall percentage of acceptable responses was a total of 64.73, for 35.27 of unacceptable answers, and the difference between the two proved to be significant ($\chi^2(1) = 19.45, p < .001$), thereby confirming that participants interpreted NAND sentences correctly. And within acceptable answers, 82.07 percent of responses corresponded to the FTTT pattern and 17.93 percent to the FFFT response, and the difference was clearly significant here too ($\chi^2(1) = 59.65, p < .001$).

The results are noteworthy for a number of reasons. Firstly, the methodology and analyses we employed allowed us to probe the semantics of language's logical connectives in a more direct manner than has been the case before. Our approach was not centred on analysing the responses of a forced choice task (*11, 14*), nor did it involve evaluating which predictors are significant at specific (and short) time windows of an eye-movement record (*31*), without explicitly linking up these two sets of data. Instead, we tracked the full time-course of participants' eye movements as they processed the sentences against a background of four potentially

relevant scenes by using a GAMM to analyse these data, thereby establishing *when* exactly differences among the four readings arise. This was specifically done in order to contrast the (implicit) record of the cognitive processes involved in comprehending the sentences with the (more explicit) responses at the end of the trials, the latter the result of having to think through the sentences' full (logical) interpretation.

Thus, we were able to ascertain that participants settle on the primary interpretation of compound sentences early on, and moreover, that they do not diverge from this interpretation during the processing of a sentence, including during the few seconds after a sentence has ended, even if the scenes on display provide compatible interpretations and more reflexive processes as to what the sentences might mean could have been elicited in such conditions. As suspected, reasoning-related processes hardly need apply during the actual, online processing of a sentence, but when participants *are* asked to evaluate the overall meaning of compound sentences, as they were in this experiment, they *do* demonstrate mastery of the relevant truth tables. The latter ability is certainly part of what reasoning with compound sentences necessitates (part, that is, of the interface between comprehension and reasoning).

More to the (central) point of the study, this framework offers a compelling way to demonstrate whether participants are able to conceptualise unverbalised connectives, and in so doing, substantiate the split between the lexicalisation of the logical connectives and the learnability of non-existent words meant to embody the meaning of unlexicalised connectives. Participants showed that they could indeed learn the NAND connective, and moreover, that they interpret non-existent NAND sentences in the same way that they interpret real, disjunctive sentences. That is, there was a clear preference for what may be regarded as the core meaning of NAND sentences in the eye-movement record, more so, in fact, than in the case of disjunction, though the other valid interpretations of NAND were also available when required, much as with (inclusive) disjunction.

This result is of course directly related to the old issue of how language and thought relate, a question as unsettled as any other in the study of cognition. Putting to one side some of the issues the field has been concerned with in relation to this topic, including whether certain thoughts necessitate specific natural languages, our evidence illustrates the long-held claim that linguistic representations do not exhaust what may be in general thought in human cognition (*32*). Or put otherwise, that the universe of the things we can entertain and think about appears to be greater than the universe of the things we can verbalise and talk about. It is certainly curious that we are capable of employing the resources of language to make some non-linguistic thinking processes explicit, as in the possibility of learning the meaning of made-up words standing for unlexicalied concepts such as the NAND connective, meanings that are not part of natural language at all, but we take it that this is in line with the equally long-held idea that what goes unsaid is not necessarily unthought, that much of thought is probably unverbalised and in fact unconscious (*33*). This for us suggests that the relationship between the capacity for natural language and the *language of thought* is more intricate than it is usually taken to be (*34*), requiring a particular experimental perspective. The present framework should bode well for research on this very issue, as it ought to be possible to apply this approach to the study of other unverbalised properties of language and cognition, thereby tracking a fuller array of what can be humanly thought.

References

1. G. Frege, *Mind* **72**, 1 (1963). First publication 1923.

2. J. A. Fodor, *The Language of Thought* (Harvard University Press, Cambridge, MA, 1975).

3. M. D. S. Braine, D. P. O'Brien, eds., *Mental Logic* (Lawrence Erlbaum, Mahwah, NJ, 1998).

4. P. N. Johnson-Laird, Y. Yang, *The Cambridge handbook of computational psychology*, R. Sun, ed. (Cambridge University Press, Cambridge, England, 2008), pp. 339–58.

5. M. D. S. Braine, B. Rumain, *Journal of Experimental Child Psychology* **31**, 46 (1981).

6. S. G. Paris, *Journal of Experimental Child Psychology* **16**, 278 (1973).

7. M. Haspelmath, *Coordinating constructions*, M. Haspelmath, ed. (John Benjamins Publishing Co., The Netherlands, 2004), pp. 3–39.

8. N. Klinedinst, D. Rothschild, *Natural language semantics* **20**, 137 (2012).

9. P. H. Grice, *Studies in the way of words* (Harvard University Press, Cambridge, Massachusetts, 1989).

10. G. Chierchia, *Linguistic Inquiry* **37**, 535 (2006).

11. D. J. Lobina, J. Demestre, J. E. García-Albea, M. Guasch, *Linguistics and Philosophy* (2021).

12. G. Gazdar, G. Pullum, *Papers from the Twelfth Regional Meeting, Chicago Linguistic Society* (1976), pp. 220–234.

13. G. Gazdar, *Pragmatics: implicature, presupposition, and logical form* (Academic Press, New York, New York, 1979).

14. C. Chevallier, *et al.*, *The Quarterly Journal of Experimental Psychology* **61**, 1741 (2008).

15. S. Mascarenhas, P. Koralus, *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, D. C. Noelle, *et al.*, eds. (Cognitive Science Society, Austin, Texas, 2015), pp. 1541–1546.

16. S. Mody, S. Carey, *Cognition* **154**, 40 (2016).

17. N. Cesana-Arlotti, *et al.*, *Science* **359**, 1263 (2018).

18. L. R. Horn, *The Square of Opposition: A General Framework for Cognition*, J.-Y. Beziau, G. Payette, eds. (Peter Lang AC, Bern, Switzerland, 2012), pp. 394–426.

19. S. Guttenplan, *The Languages of Logic* (Blackwell Publishing, Oxford, England, 1997).

20. L. R. Horn, *A natural history of negation* (CSLI Publications, Standord, California, 1989).

21. R. Katzir, R. Singh, *Linguistics and philosophy* **36**, 1 (2013).

22. J. Collins, *Mind and Language* **26**, 234 (2011).

23. J. A. Fodor, E. Lepore, *The compositionality papers* (Oxford University Press, Oxford, England, 2002).

24. T. Hunter, J. Lidz, A. Wellwood, A. Conroy, *Proceedings of Semantics and Linguistic Theory (SALT) 19*, E. Cormany, S. Ito, D. Lutz, eds. (eLanguage, 2011), pp. 223–238.

25. R. H. Baayen, J. van Rij, C. de Cat, S. Wood, *Mixed-Effects Regression Models in Linguistics*, D. Speelman, K. Heylen, D. Geeraerts, eds. (Springer International Publishing AG, Cham, Switzerland, 2018), pp. 49–70.

26. D. Mirman, J. A. Dixon, J. S. Magnuson, *Journal of Memory and Language* **59**, 475 (2008).

27. D. J. Barr, *Journal of Memory and Language* **59**, 457 (2008).

28. S. N. Wood, *Generalized Additive Models: An Introduction with R* (Chapman and Hall/CRC, London, England, 2017), second edn.

29. R. H. Baayen, S. Vasishth, R. Kliegl, D. Bates, *Journal of Memory and Language* **94**, 206 (2017).

30. G. Montero-Melis, T. F. Jaeger, *Bilingualism: Language and Cognition* **23**, 602 (2019).

31. L. Zhan, *Frontiers in Psychology* **9**, 1 (2018).

32. C. Panaccio, *Mental Language* (Fordham University Press, New York, NY, 2017).

33. J. A. Fodor, *London Review of Books* **13** (1991).

34. D. J. Lobina, *Linguistic and Philosophical Investigations* **18**, 37 (2019).

Acknowledgments

Supplementary materials

Supplementary Text

Materials and Methods

Figs. S1 to S6

Tables S1 to S5

References *(35–49)*

A

i

The figures will be triangles and squares

The triangle is yellow OR the square is blue
The triangle is yellow NAND the square is blue

Eye Tracking

ii

Behavioural Response

iii

B

Disjunction Experiment

NAND Experiment

C

Disjunction Experiment		
Responses	Frequency	Percent
TTTF	122	50.83
FTTF	29	12.08
TFFF	68	28.33
Other	21	8.76

NAND Experiment		
Responses	Frequency	Percent
FTTT	119	49.58
FFFT	26	10.83
Other	95	39.59

Figure 1: Interpreting and thinking with existent and non-existent verbal connectives. **(A)** (i) Panel specifies the figures participants will see in the trial. (ii) Each quadrant in the experimental panel presents two figures in diverse combinations of colours. An audio of a sentence describing one or more quadrants is played soon after the figures appear, as shown to the right of the panel. Each quadrant represents a combination of the truth values of two propositions, as in the two first columns of Table 1 (namely, the combinations TT, TF, FT, and FF), while the sentence played to participants is a linguistic representation of two propositions ("clauses", in language) mediated by a logical connective (a linguistic coordinator). The particular linguistic connective used in each experiment signals a specific interpretation —i.e., a unique truth table, as shown in columns 3–7 of Table 1 (e.g., TFFF is the truth table, or meaning, of the connective *conjunction*). In Experiment 1 participants would hear a disjunctive sentence such as *the triangle is yellow OR the square is blue*, while in Experiment 2 participants would first undertake a learning session to expose them to the meaning of a made-up word standing for the unlexicalised connective NAND and then in the experimental session they would hear sentences such as *the triangle is yellow NAND the square is blue* (see **Materials and Methods** for details). The tracking of eye movements starts from the beginning of the sentence and continues for a further 3 seconds after the sentence finishes. (iii) At the end of the eye-tracking, the mouse pointer is activated and participants are asked to select all the quadrants that match the sentence. **(B)** Proportion of fixations to each quadrant (TT, TF, FT, and FF) for a duration of circa 6500 milliseconds (3500 ms for the longest sentence plus 3000 ms of 'looking time'). In Experiment 1 (left panel), participants mostly fixated on the TT and TF quadrants after c. 1000 ms and on mostly TT only after 3000 ms. In Experiment 2 (right panel), participants fixated almost exclusively on the FF quadrant from c. 2000 ms. **(C)** Behavioural responses, where the TTTF pattern, for instance, specifies that participants had selected quadrants TT, TF, and FT (for reference, see Table 1). In Experiment 1 (table on the left), participants selected

22

the inclusive interpretation of disjunction (TTTF) more often than the exclusive interpretation (FTTF) or a 'conjunctive' reading (the TFFF pattern). In Experiment 2 (table on the right), participants selected the FTTT pattern, the truth table of NAND, as the best interpretation for the non-existent verbal connective, in preference to the FFFT pattern, the truth table of NOR, another unlexicalised connective.
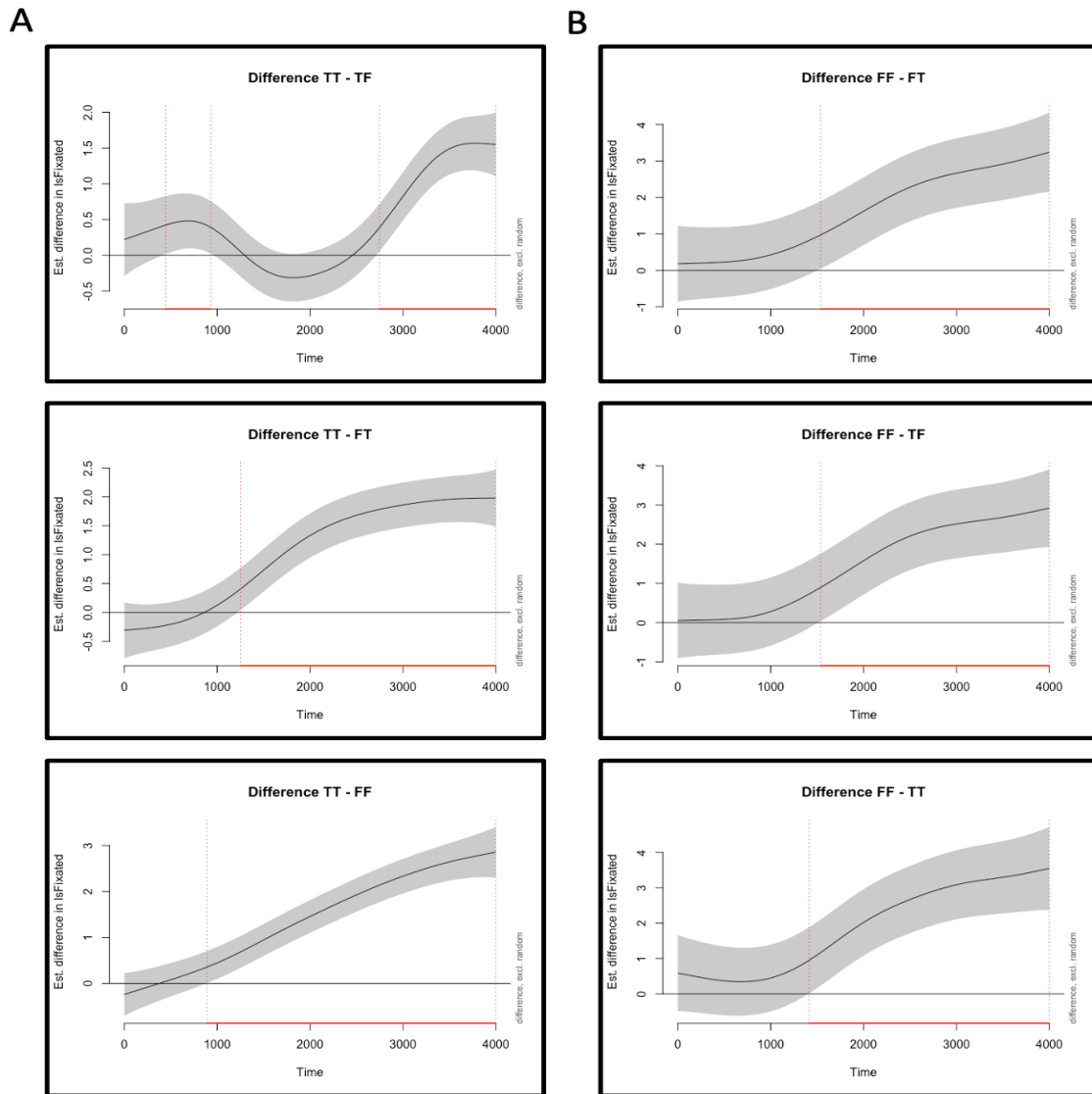
Figure 2: Difference curves, derived from the best-fit model of a generalised additive mixed-model analysis. The graphs show the comparison between the (non-linear) smooth of the quadrant with the most fixations (fixations are labelled as IsFixated on the y-axis) against each of the (non-linear) smooths of the other quadrants, with the gray solid line indicating the estimated difference. The shaded band represents the pointwise 95%-confidence interval; when the band doesn't overlap with the x-axis (i.e., the value is significantly different from zero), this is in-

24

dicated by a red solid line on the x-axis along with red vertical dotted lines. The graphs show fixations during the audio of the entire sentence in addition to an extra 640–1000 ms of "looking time" (see **Materials and Methods** for the selection of this time-window for the analyses). **(A)** Estimated differences from Experiment 1 between fixations to the TT quadrant and fixations to the TF, FT, and FF quadrants. The TT-TF comparison exhibits differences in two time windows, at window 445–930 ms, roughly around the time the first clause is being played in the audio, and at 2750–3999 ms, where the beginning of this time-window coincides with the end of the audio. The TT-FT contrast produces a difference at 1250–3999 ms, a window that (roughly) starts right after the connective has been presented, while the TT-FF comparison exhibits a difference at 890–3999 ms, where the beginning of this time-window precedes the presentation of the connective. **(B)** Estimated differences in Experiment 2 between fixations to the FF quadrant and fixations to the FT, TF, and TT quadrants. Comparisons show that all differences surfaced after around 1400 ms, when the connective has already appeared and the second clause is being presented.

Supplementary Materials for
Unsaid Thoughts: Thinking in the absence of (some) verbal logical connectives

David J. Lobina,[1]* Josep Demestre,[2] José E. García-Albea[2] & Marc Guasch[2]

[1]Psycholinguistics Research Group, Universitat Rovira i Virgili

[2]Department of Psychology and Research Center for Behavior Assessment,
Universitat Rovira i Virgili, Ctra. de Valls s/n, 43007 Tarragona, Spain

[3]Department of Psychology, Universitat Rovira i Virgili,
Ctra. de Valls s/n, 43007 Tarragona, Spain

*To whom correspondence should be addressed; E-mail: djlobina@gmail.com

Contents

1   A note on analysing eye-movement data

The analysis of eye-movement data can present some challenges in the study of cognition and there is at present no consensus on how best to approach this issue. As is usually the case, much depends on the actual design of an experiment. A study employing the *visual world paradigm* typically involves measuring participants' fixations to specific regions of a visual display given a specific stimulus; in a psycholinguistic study such as this one, the input stimulus is often a sentence describing some aspect of the scene that is presented to participants on a computer screen. In the case of a display showcasing 4 objects or graphics, the design used here, the screen can be divided into quadrants, or areas of interest (AOIs), where each graphic is placed on a quadrant and each AOI is classified according to how each graphic relates to the sentence played during the trial —viz., as the target interpretation for the sentence, as a competitor for the targeted interpretation, as a distractor, etc. (in most cases, a 4-way display exhibits one target, one competitor, and two distractors). In such a setting, the analysis is usually centred on whether the AOIs differ significantly on the likelihood of being fixated on, including when any such differences arise during the time course of a trial, with the dependent variable commonly analysed as "looks to the target AOI versus looks elsewhere" (where "looks" can be treated as proportions of fixations, log odds, binomial data, etc.). A number of techniques have been employed to inferentially test observed differences, but not without problems.

A widely-used way to analyse these data in the past was to compare proportions of fixations to different AOIs on specific time windows by using ANOVAs or t-tests (proportions of fixations are means of raw proportion scores, with a range between '0' and '1'). However, the use of ANOVAs and t-tests on time-series data can often violate certain assumptions of these statistical methods (*35*), and even when this is not the case, the comparisons may only unearth rather general patterns (e.g., whether a given comparison is significant in a time window of 500

2

milliseconds, for instance, may not be very informative). A better way to analyse eye-tracking data is to convert proportions of looks to binomial data —a fixation on a given AOI at a particular time would be coded as '1', or success, and '0' otherwise— and to employ a generalized linear mixed-effect model to run a logistic regression (*35, 36*). This can be complemented with orthogonal polynomials time functions to implement a so-called *growth curve analysis* (*37*), which allows researchers to assess the predictive value of the experimental conditions over time (that is, the analysis can tell you which time polynomials interact with the conditions under investigation).

The design of our experiments presented additional analytical challenges and a *growth curve analysis* was not entirely appropriate for the purposes of the study; among other reasons, this kind of analysis does not reveal the precise moment in time where differences between AOIs become significant, and this is what we were actually interested in. As explained in the main text, in our experiments each AOI exhibited a unique combination of the truth values of the truth table of a logical connective (4 combinations to a truth table: True-True, True-False, False-True, and False-False) and thus each AOI was potentially relevant for the interpretation of the compound sentences participants would be exposed to. In addition, the length of the eye-tracking record of our experiments, a total of 6000 milliseconds per trial (though see below for a clarification), was likely to give rise to two problematic issues: that over such a long record fixations might exhibit a non-linear pattern, as indeed suggested by the wiggliness of the lines representing proportions of fixations shown in Figure 1, Panel B, a plot that was generated to visually inspect the data before analysis; and that there would be a significant amount of autocorrelation in the data, a common occurrence in experiments dealing with time-series (*38*).

The autocorrelation issue was especially problematic in our study on account of some features of the experiments. In particular, the overall task was effectively underlain by a four-way, multinomial choice, as all AOIs were technically relevant and thus eye movements to all four

AOIs needed to be tracked. In order to account for this aspect of the task, we recoded the probabilities of fixations to each AOI as binomial data, where '1' would mark a fixation on a specific AOI at a particular time, with the remaining three areas coded as '0' as non-fixations at that same time ("area of interest" was one of the predicting factors included in the analyses and constituted the indicator variable for the coding of the multinomial model). Crucially, this meant that the data had to be recoded on a millisecond by millisecond basis and thus at the sampling rate the eye tracker recorded the eye movements, whereas in most visual world studies the data are downsampled to 20 millisecond chunks prior to the analyses, which can decrease the amount of autocorrelation significantly, though not completely, and some information (or patterns) may be missed by so doing. In the case of our experiments, the amount of autocorrelation was bound to be very high and this was indeed confirmed in the analyses, as explained below.

We employed a *generalised additive mixed-model* (GAMM) (*39*) to account for these issues, a statistical technique that is becoming more common in analyses of time-series data (*38, 40*), including a few eye-tracking studies (*41*). GAMMs are especially useful on account of three features: the models relax the assumption of a linear relationship between predicting variables and response variables by implementing *smooth functions*, including so-called *factor smooths*, which can implement wiggly random effects (the non-linear equivalent of random intercepts and random slopes); autocorrelation can be accounted for by the inclusion of an autoregressive AR-1 parameter (in combination with factor smooths); and, particularly convenient for this study, the interpretation of the models is partly determined visually, allowing us to plot, among other things, the differences between conditions at specific moments in time.

The visual evaluation of GAMMs is very important and constitutes one of the three ways to test the significance of these models. The other two involve inspecting the summary statistics of each model and conducting model comparisons in terms of chi-square tests on both fREML scores and the difference in degrees of freedom specified in each model (model comparison

4

in terms of AIC scores, the usual way to compare linear models is not reliable when an AR-1 parameter is included). We shall employ all three methods in the analyses below.

2   Materials and Methods

Given that the present study was centred on the semantic and logical properties of compound sentences and the core objective was to probe the availability of the unlexicalised connective NAND, a number of issues required special attention. First of all, we used experimental materials that minimised some of the pragmatic effects that sometimes arise when interpreting such sentences, in line with a framework employed before to this effect (*42*). This was potentially a factor in Experiment 1, where we used disjunctive sentences as the experimental condition and such sentences tend to elicit an exclusivity implicature, thereby countenancing the inclusive interpretation of disjunction. It is unclear whether similar effects would surface in Experiment 2, where we employed compound sentences in which the two clauses were connected by a made-up word standing for the NAND connective. In principle, participants have no experience whatsoever with NAND sentences, and barring any problems regarding their ability to learn the meaning of a non-existent word for NAND during the learning phase of this experiment, no further, extraneous effects were expected or predicted.

The sentences we designed simply ascribed different colours to various geometrical figures, thus greatly restricting the context participants are offered as well as being rather neutral as to what interpretation would be favoured. An example of each condition from each experiment is presented below in Spanish, the language of the study (the translations appear in parentheses; note that in Experiment 2 we used the non-existent but possible Spanish word *fro* for the NAND connective; see below for more details).

Or condition, Experiment 1: El círculo es azul *o* el cuadrado es amarillo (the circle is blue OR
    the square is yellow)

NAND condition, Experiment 2: El círculo es azul *fro* el cuadrado es amarillo (the circle is

     blue NAND the square is yellow)

In addition, we took two different kinds of measurements in each trial —the participants' eye-movement record as the sentence was played, and behavioural responses at the end of each trial when participants were asked to select all the quadrants they thought matched the sentence they had heard— and it was hoped that in combination these data would provide a more comprehensive picture of the overall interpretation of the manipulated sentences. Participants might exhibit a preference for an exclusive interpretation of disjunction in their responses, for instance, but the eye-tracking data could potentially show what sort of (implicit) consideration they give to relevant other readings, and this was deemed to be of potential interest for NAND sentences as well (as proved to be the case, as shown below and discussed in the main text).

Finally, and as a way to make sure performance on the NAND condition would not be affected by the nature, or novelty, of the general set-up, participants were required to carry out both experiments in the same session (with various breaks), first Experiment 1 and then Experiment 2. It was thought that Experiment 1, which did not employ any non-existent words but language's disjunction *or*, would provide a gentle sort of introduction to the overall setting so that participants were sufficiently familiar to the task at hand by the time they were exposed to NAND sentences in Experiment 2. Further, disjunctive sentences were specifically chosen in order to avoid the contrast participants might have considered between the NAND connective and its contrary, conjunction *and* (see Table 1), which could have introduced an unfortunate artefact. Namely, by using conjunctive sentences participants might have realised that correctly interpreting NAND sentences simply required entertaining the opposite set of facts to conjunction, but no such comparison was possible between *or* and NAND (though there was an interesting point of contact between the two in relation to the unlexicalised connective NOR, the contrary of inclusive disjunction; we come back to this below).

2.1 Experiment 1

Participants. 15 psychology students (1 male, 14 female) from the Rovira i Virgili University (Tarragona, Spain) participated in this and the next experiment for course credit. None of the participants had undertaken any course in logic or reasoning before taking part in the experiments. All participants carried out Experiment 1 first, and after a short break, Experiment 2. The mean age was 18.5 years ($SD = 0.68$), and none of the participants had any known hearing or visual impairments. All were native speakers of Spanish. All participants gave their written informed consent before taking the experiment.

Materials. This experiment evaluated a single condition with 4 levels. Each condition corresponds to each of the four values of the inclusive disjunction's truth table (TTTF). 16 biclausal, declarative Spanish sentences, with the two clauses connected by the coordinator *or*, were constructed. Each clause ascribed a single colour to a single geometrical figure; we used four different colours (blue, yellow, red, and green) and four different figures (circles, squares, triangles, and diamonds). 16 more sentences were constructed to act as fillers. The fillers were monoclausal and thus only one figure was mentioned, but in this case the figure was ascribed two colours instead of one by employing the connective *and* and the overall sentence was furthermore negated (e.g., *the circle is not blue and green*). A further 8 practice sentences were created, four of which were similar to the experimental items and four to the fillers. The average length of all sentences was 2555 milliseconds (ms) and the longest sentence was around 3000 ms. The sentences were recorded in stereo with a normal but subdued intonation by a native, male speaker of the Spanish language using the Praat software on a Windows-operated computer. The graphics representing each one of the truth values of inclusive disjunction as well as the truth values of negated conjunctions (for the fillers) were created with Microsoft PowerPoint.

Procedure. The experiment was designed and run with the Experimental Builder software (SR

Research Ltd.) and administered in a laboratory with low to normal illumination in which each participant was tested individually. Participants were seated in front of a computer screen and were asked to place their head on a chin rest. The chin support was adjusted for each participant so that there was a distance of around 60 centimetres between their eyes and the monitor where the visual scene was presented, a 19-inch screen set to a resolution of 1024×768 pixels. The position and fixations of participants' right eye, most people's dominant eye, were continuously recorded at a sampling rate of 1000 Hz with an EyeLink 1000 eye tracker. In addition to the eye-tracking data, the participants' behavioural responses at the end of the trials were also recorded.

The overall flow of the experiment as well as the general design is shown in Figure 1, Panel A. Each trial started with a fixation point in the middle of a white screen. Participants were asked to fixate on this point and to press the space bar when they were ready to start the trial. A sentence such as *the figures will be triangles and squares* would replace the fixation point 500 ms after pressing the space bar. The sentence would stay on the screen for 2500 ms so that participants had enough time to read it fully; at the end of this period of time, the sentence was replaced by the visual display, which would remain on the screen for the remainder of the trial. The display was divided into quadrants and a specific combination of figures and colours would appear on each quadrant, where the figures matched those announced in the sentence presented before the visual display (the placement of each graphic was randomised across quadrants and trials). After 2000 ms, the time we allocated to participants to view the quadrants fully before presenting any other stimuli, a sentence describing one or more quadrants was played over headphones binaurally. Once the sentence had finished, there was a period of 3000 ms of "looking time", at the end of which the cursor would be activated so that participants could select the quadrants they thought the sentence described appropriately. The trial ended once participants were satisfied with their answers and had pressed the space bar to move on to the next trial (or reach the end of the session). Participants carried out an 8-item practice session with the

8

experimenter, who explained the overall task and answered any questions before proceeding to the experimental session. Eye calibration was conducted before the practice session and again before the experimental session. The experimental session consisted of a total of 32 items, 16 of which were experimental and 16 fillers; the presentation of experimental and filler sentences was randomised. The experiment lasted 15 minutes overall.

Results. *Eye-tracking data*. The eye gaze data collected with the EyeLink 1000 eye tracker was exported by using the manufacturer's Data Viewer software. The Sample Report this software outputs requires significant preprocessing before analysis and plotting, and we used the R package *VWPre*, version 1.2.4, for this purpose (*43*). To begin with, we performed an analysis of trackloss (i.e., the amount of times the eye tracker lost track of participants' eye gaze); 1.68% of data was marked as off-screen and 4.86% as trackloss, and as a result 7 trials with less than 75% of data were eliminated (this threshold is common in the literature and seemed reasonable for our own experiments too). The data were then prepared in order to conduct a logistic GAMM analysis. As the task was effectively underlain by a four-way, multinomial choice and thus all AOIs were potentially valid interpretations for the sentences, the factor "area of interest" (AOI) was the main predictor in our models and constituted the indicator variable for the coding. The data were coded on a millisecond by millisecond basis, where a fixation on a given AOI at a given time was coded as a '1' (for success) and a non-fixation as a '0' (for failure).

The logistic GAMMs were conducted and partly analysed with the R package *mgcv*, version 1.8-35, family class binomial (*39*). Three such models were run and compared, and their analysis was complemented with the R package *itsadug*, version 2.4 (*44*). We used model selection to determine the best random-effects structure for the data, which is the most appropriate approach for non-linear models such as GAMMs (*40, 45*). The R script for the analyses of this and the next experiment as well as the eye-tracking data can be found at: `https://osf.io/mfqt8/?view_only=dc416b5ac605423b80449714fd0f4979`.

9

An important factor to consider when using the *mgcv* package is that fitting GAMMs often requires significant computational resources and some processes may take a very long time to complete, in some cases even days (fitting such models certainly takes much longer than fitting linear mixed-effects models with the *lme4* package, a staple of contemporary research). Given that in the 6000 ms of eye movements we recorded in each trial a pattern was pretty evident after 4000 ms and did not change in any way after that, as shown in Figure 1, Panel B for both experiments, we decided to reduce the length of the time-series to analyse to 4000 ms, which was hoped would ease the demands of the analyses. The average length of the experimental sentences from Experiment 1 was 2947 ms and this meant that the models we analysed included at least 1000 ms of "looking time", which was abundantly sufficient for the purposes at hand and thus no important effects were expected to be missed.

Three models were run, one of which, M0, was unlikely to fit the data very well and was primarily used to estimate the amount of autocorrelation to correct in the models we expected to fit the data better, models M1 and M2. In all models, the AOI variable was fitted to the response variable IsFixated, which as discussed earlier was treated as a binary kind of data ('1' for a fixation on an AOI at a given time, '0' otherwise). All three models included a smooth function for Time by AOI (TT, TF, FT, and FF; the function appears as *s(Time)* in the model itself, as shown in Table 3), allowing us to assess whether there was a non-linear relationship between the fixed-effect AOI condition and the response variable IsFixated over time (that is, this smooth function assessed the one-dimensional —there is only one numerical predictor— interaction between the factors Time and AOI). Regarding the random effects, model M0 was fitted with by-participants and by-items random intercepts and slopes (4 curves in total), which were not expected to capture the variability very well given the potential non-linearity of the data. A visual inspection of the autocorrelation, shown in Figure 2, Top Panel, confirmed the high amount of autocorrelated data and furthermore indicated that there was only a slight

decrease in such autocorrelation across the time course of trials (the decrease is more marked in downsampled, 20 ms chunks (*46*)). Following the general practice from the literature, we took the amount of autocorrelation at time lag 1, $\rho = 0.995$, to be an appropriate measure for the amount of autocorrelation to be corrected (*38*). M0 was rerun with an AR-1 parameter and the autocorrelation was eliminated, as shown in Figure 2, Bottom Panel (this was the case for every other model we ran and thus abstain from adding any more graphics on this issue). The summary statistics of model M0 indicated that its random structure did not account for much of the variability (only one of the four curves was statistically significant) and thus we disregarded this random structure from subsequent models (further, a comparison of the summary statistics of a model that included an AR-1 parameter with a model that did not include this parameter illustrates the unreliability of models that do not control for autocorrelated data; these summary statistics are not included here, however).

Models M1 and M2 were set up to account for non-linear random effects, though in different ways, as we shall see. What these models did share, as mentioned, was a smooth function for Time by AOI as well as an AR-1 parameter to correct the autocorrelation, and to this were added functions to model possible trial effects, one of the "human factors" the field is ever more concerned with (*40*). In particular, we added a smooth function for Trial by AOI (*s(Trial)* in the model) to assess the (one-dimensional) interaction between trial number and AOI, and a tensor product to evaluate the two-dimensional interaction between the factors time and trial number by AOI (this interaction is two-dimensional because there are two numerical predictors to compare, one for the Time factor and one for Trial Number factor; the tensor product itself appears as *ti(Time, Trial)* in the model). As for the random effects, M1 was fitted with by-participant and by-item factor smooths for Time by AOI, replacing the random intercepts and slopes from M0 and added to this model in order to implement wiggly random effects (factor smooths are centred and in fact penalised against non-linearity), while M2 included factor smooths for Time

11

per time-series by AOI (i.e., per trial; *s(Time, Event)* in the model), replacing the factor smooths from M1 and potentially offering a closer fit to the data ("event" smooths tell the model that the measurements for each time series are not in fact independent, a feature that ought to decrease the amount of autocorrelation as well). All together, model M1 yielded 20 parameters (or curves) and model M2 a total of 16. In order to evaluate these models, we start with the summary statistics, followed by model comparison and then, crucially, as we shall point out, a visual evaluation of the best-fit model.

For every smooth function of M1 and M2, except for the smooth for Trial, non-linear curves were obtained, as confirmed by the effective degrees of freedom (edf), a summary statistic of GAMMs that reflects the degree of non-linearity of a curve. An edf equal or close to 1 corresponds to a linear relationship between the predicting variables and the response variable, and anything above 2 equals to a highly non-linear relationship, which is what was observed in the two models for every smooth except, as alluded to, for the effect of Trial by AOI (with a $p < .001$ in every case a non-linear relationship was observed). As for the Time x Trial interaction, there was a significant effect regarding the TT AOI ($p = .044$ in M1, $p = .028$ in M2). In such circumstances, the predictor Trial by AOI would normally be eliminated from the models, but a GAMM that includes a tensor product to evaluate a two-dimensional interaction assumes that there is a smooth function for the first factor of the interaction (here, Time) and a smooth function for the second factor (here, Trial Number), and therefore the smooth for Trial by AOI was kept in the final analyses (GAMMs also allow modelling a type of tensor product that implicitly includes the smooths for each of the two factors being compared, a so-called *te*-operator, but this would not have been conducive to our objectives, as we needed to directly asses the significance of the smooth function for Time by AOI, the fixed effect).

Regarding model comparison, we used a Maximum Likelihood (ML) score comparison, which is preferred over a comparison in terms of AIC scores, as AICs become unreliable (i.e.,

anti-conservative) when autocorrelation is included in a model. Table 2 shows the details of the comparison between M1 and M2, where the differences in fREML scores and degrees of freedom indicate a preference for model M2, the model with event factor smooths instead of by-participants and by-items factor smooths. Table 3, in turn, provides the full summary statistics of the best-fit model, M2 (note the values of the various edfs as well as the p-values for each one of the non-linear curves).

As for the visual inspection of the best-fit model, three graphics are relevant in this case. Figure 3 shows the fixations on each quadrant according to M2, and a useful comparison can be drawn between the response variable IsFixated from this graphic to the proportions of fixations from Figure 1, Panel B. Two-dimensional interactions also require visual evaluation in GAMMs; Figure 4 provides a contour surface of the predicted fixations on the TT quadrant for the interaction between Time and Trial Number. As can be seen in the graphic, the likelihood that the TT quadrant is fixated on generally increases in time during each trial, but at 4000 ms fixations on the TT quadrant are less likely as the experimental session progresses —that is, the likelihood that the TT quadrant is fixated on actually decreases from trial 20 onwards towards the end of the time series, as indicated by trial number on the y-axis (these numbers refer to all items, but the data from the graphic only includes the fixations on the experimental items). This effect is rather significant given that model M2 included event smooths, which could have captured many of the effects to do with trial number.

In turn, and rather importantly, Figure 6, Panel A, shows the "difference curves" from M2 between the smooth function for TT, the most fixated AOI, against the remaining three smooths (TF, FT, FF), where the relevant time windows are marked —that is, the graphics show the estimated differences between fixations to the TT AOI and fixations to the TF, FT, and FF AOIs. These curves as well as the accompanied statistics are implemented by the package *itsadug* and constitute the most relevant way to assess whether the different levels of the AOI

13

condition significantly differ from each other in this experiment (GAMMs allow other ways to probe whether the different levels of an experimental condition differ from each other, such as by employing ordered or binary factors, but we put this issue to one side here). In particular, the graphs from Figure 6 show the comparison between the (non-linear) smooth of the AOI with the most fixations (fixations are labelled as IsFixated on the y-axis) against each of the (non-linear) smooths of the other AOIs, with the gray solid line indicating the estimated difference. The shaded band represents the pointwise 95%-confidence interval; when the band doesn't overlap with the x-axis (i.e., the value is significantly different from zero), this is indicated by a red solid line on the x-axis along with red vertical dotted lines.

The graphs show fixations during the audio of the entire sentence in addition to an extra 1000 ms of looking time. The TT-TF comparison exhibits differences in two time windows, at 445–930 ms, around the time the first clause is being played in the audio, and at 2750–3999 ms, where the beginning of this time-window coincides with the end of the audio. The TT-FT contrast produces a difference at 1250–3999 ms, a window that (roughly) starts right after the connective has been presented, while the TT-FF comparison exhibits a difference at 890–3999 ms, where the beginning of this time-window precedes the presentation of the connective.

Behavioural responses. Regarding the behavioural data, presented in Figure 1, Panel C, the preferred pattern of response for disjunction *or*, shown on the table on the left-hand side, was TTTF —that is, participants selected the TT, TF, and FT quadrants— in line with the truth table of logic's inclusive disjunction. For the analysis of these data we drew a distinction between acceptable (or correct) patterns of response and unacceptable (or incorrect) patterns in order to run chi-squared tests between the expected and observed responses, in two steps: first between acceptable and unacceptable responses as a way to confirm that the sentences had been interpreted correctly, and then within acceptable responses between the two patterns of interpretation we had identified beforehand as being correct. In this experiment, we took the

patterns TTTF and FTTF to be the only acceptable responses, as these constitute the truth tables of inclusive and exclusive disjunction, respectively, and thus everything else was regarded as a mistake in interpretation. The percentage of correct answers amounted to 67.41, for 32.59 of incorrect answers, and the difference between these two frequencies was significant ($\chi^2(1) = 27.16, p < .001$). Once it was ascertained that the task had been appropriately carried out by the participants, we compared the two patterns of acceptable responses. In this case, the percentage of responses for the inclusive interpretation of disjunction (TTTF) was 80.79, for 19.21 for the exclusive interpretation (FTTF), and this difference was clearly significant as well ($\chi^2(1) = 57.29, p < .001$).

2.2 Experiment 2

Participants. The same as in Experiment 1, as noted above.

Materials. This experiment also evaluated a single experimental condition with 4 levels, but in this case the two clauses of the compound sentences were connected by a non-existent but possible word in Spanish standing for the unlexicalised logical connective NAND (truth table: FTTT): *fro*. The nonsense word *fro* is not related to, nor does it resemble, any other word in Spanish (or Catalan, the experiment took place in Catalonia), it is easy to pronounce, and its morphology favours the sort of use participants would be exposed to in the experiment (i.e., as a coordinator). We used the same figures and colours employed in Experiment 1, and a total of 32 NAND sentences were constructed, 16 experimental sentences and 16 sentences meant for the learning phase. As for the filler sentences, these were also similar to those of Experiment 1, but in this case the monoclausal sentences were not negated. A further 8 practice sentences were created, four of which were similar to the experimental items and four to the fillers. The average length of all sentences was 2837 ms and the longest sentence was around 3500 ms. All other details remained the same as in Experiment 1, except that that the graphics for the experimental

15

sentences now represented each one of the values of the truth table for NAND and the graphics for the fillers the values of the truth table for (non-negated) conjunctive sentences.

Procedure. The procedure was very similar to that of Experiment 1, with the addition of a learning phase for the connective NAND. In this phase, which was undertaken before the practice session, participants would be shown a series of situations that NAND sentences could appropriately describe (or not). In particular, participants would be exposed to individual graphics in each trial, with each graphic always depicting two geometrical figures in two different colours, as in the experimental materials. The rationale was that participants would be exposed to individual graphics in each trial so that each one of the truth values from the truth table for NAND would be presented individually. Regarding the actual procedure of the learning phase, the graphic would appear on the screen first, and after a brief period of time for participants to inspect it adequately, a NAND sentence would be played over the headphones. Soon after participants would be presented with a feedback on screen indicating whether the sentence was an appropriate description of the graphic or not —with a tick, for 'yes', and an X for 'no'. Participants undertook four iterations of the truth table for NAND, for a total of 16 trials; that is, the four values of the truth table of NAND —viz., FTTT— were repeated four times so that participants were exposed to 16 different NAND sentences (the order of presentation was randomised). There was no explicit instruction of any kind, nor did participants have to complete any task; instead, the learning phase was similar to what is employed in artificial grammar learning studies (*47*) and the expectation was that participants would implicitly learn the meaning of the novel word being presented to them given a specific set of scenes (a similar strategy is usually employed in studies of language acquisition (*48*)). Once this phase had been completed, participants undertook an 8-item practice session and right after an experimental session of 32 items (16 experimental sentences, 16 filler sentences); all other aspects of the experiment remained the same as in Experiment 1 (4 areas of interest per trial, randomisation, etc.). Eye

16

calibration was conducted before the practice session and again before the experimental session, but it was not included in the learning phase. This experiment lasted around 20 minutes overall.

Results. *Eye-tracking data:* The eye gaze data was prepared in the same way as the data from Experiment 1 was prepared. In this case, an analysis of trackloss marked 1.83% of data as off-screen and 6.08% as actual trackloss, and thus 11 trials with less than 75% of data were eliminated. In this experiment too the time series were reduced to 4000 ms and this yielded around 640 ms of looking time (the average length of experimental sentences was 3366 ms), which was also amply sufficient for the analyses, as the overall pattern of this experiment was in fact established significantly early, around 1400 ms. GAMMs were processed and analysed in the same way as in Experiment 1, and the same three kinds of models were run —that is, the details of models M0, M1, and M2 were exactly like the corresponding models from Experiment 1. In this experiment, the amount of autocorrelated data was estimated at $\rho = 0.996$ after running the equivalent M0 model, and in this case too this model was discarded because its random structure failed to capture the data adequately.

Models M1 and M2 from this experiment were evaluated in the same way that the models from Experiment 1 had been evaluated. In terms of summary statistics, there were hardly any trial effects in either model, in terms of either the smooth function for Trial by AOI (expect for one borderline case; see Table 5) or the tensor product for the Time x Trial interaction. In all other cases, the smooth functions resulted in non-linear curves in every case, with all edfs above 2 and $ps < .001$. Model comparison favoured model M1 over model M2, as shown in Table 4, where the chi-square statistic indicates a significant difference in likelihood between the models, given fREML scores and the degrees of freedom, and the p-value suggests that there is strong evidence in favour of model M1. That is, model M1, which included by-participants and by-items factor smooths —*s(Time, Subject)* and *s(Time, Item)* in the model from Table 5, respectively— was preferred to model M2, which included event smooths. Thus, and though

a smooth for Time per time-series (per event) is usually expected to account for the random structure better than the by-participants and by-items factor smooths, as had been the case for the best-fit GAMM for disjunction *or*, this did not prove to be so in this case.

Regarding the visual inspection of model M1, two graphics are relevant. Figure 5 shows the fixations on each quadrant according to M1, and once again it may be useful to compare the IsFixated, binomial data from this graphic to the proportions of fixations from Figure 1, Panel B. Figure 6, Panel B, in turn, shows the model's difference curves, or estimated differences, between fixations to the FF AOI and fixations to the FT, TF, and TT AOIs. Comparisons show that all differences surfaced after around 1400 ms, when the connective NAND had already appeared and the second clause of the sentences was being played.

Behavioural responses: The analysis of the behavioural data of this experiment presented a slightly more nuanced state of affairs than had been the case in Experiment 1. Whilst the full interpretation of the NAND connective would correspond to the FTTT pattern —that is, the quadrants TF, FT, and FF, which was the most frequent response, in fact— the FFFT response (that is, only quadrant FF is true), properly speaking the truth table of the (unlexicalised) connective NOR, the contrary of inclusive disjunction, cannot be regarded as an entirely incorrect or unacceptable response for NAND, but perhaps simply an incomplete one. As noted, NAND is the opposite of conjunction, and in this sense, the FF interpretation is plausibly the most straightforward counterpart to the TT reading of conjunction —or the default, preferred interpretation, as the eye-movement data actually seem to indicate. Indeed, participants overwhelmingly fixate on the FF quadrant, which on its own may have indicated that participants understood the NAND connective as actually the NOR connective (as is the case in the relationship between inclusive and exclusive disjunctions, the true values of NOR's truth table constitute a subset of the true values of NAND's truth table). As such, we included FFFT as the second acceptable response along with FTTT. Thus established, the overall percentage of acceptable responses was a total

18

of 64.73, for 35.27 of unacceptable answers, and the difference between the two frequencies proved to be significant ($\chi^2(1) = 19.45, p < .001$), confirming that NAND sentences were interpreted correctly. And within acceptable answers only, 82.07 percent of responses corresponded to the FTTT pattern and 17.93 percent to the FFFT response, and the difference was clearly significant here too ($\chi^2(1) = 59.65, p < .001$).

3 Figures and Tables

3.1 Figures

A

i

The figures will be triangles and squares

The triangle is yellow OR the square is blue
The triangle is yellow NAND the square is blue

ii

Eye Tracking

iii

Behavioural Response

B

Disjunction Experiment

NAND Experiment

C

Disjunction Experiment		
Responses	Frequency	Percent
TTTF	122	50.83
FTTF	29	12.08
TFFF	68	28.33
Other	21	8.76

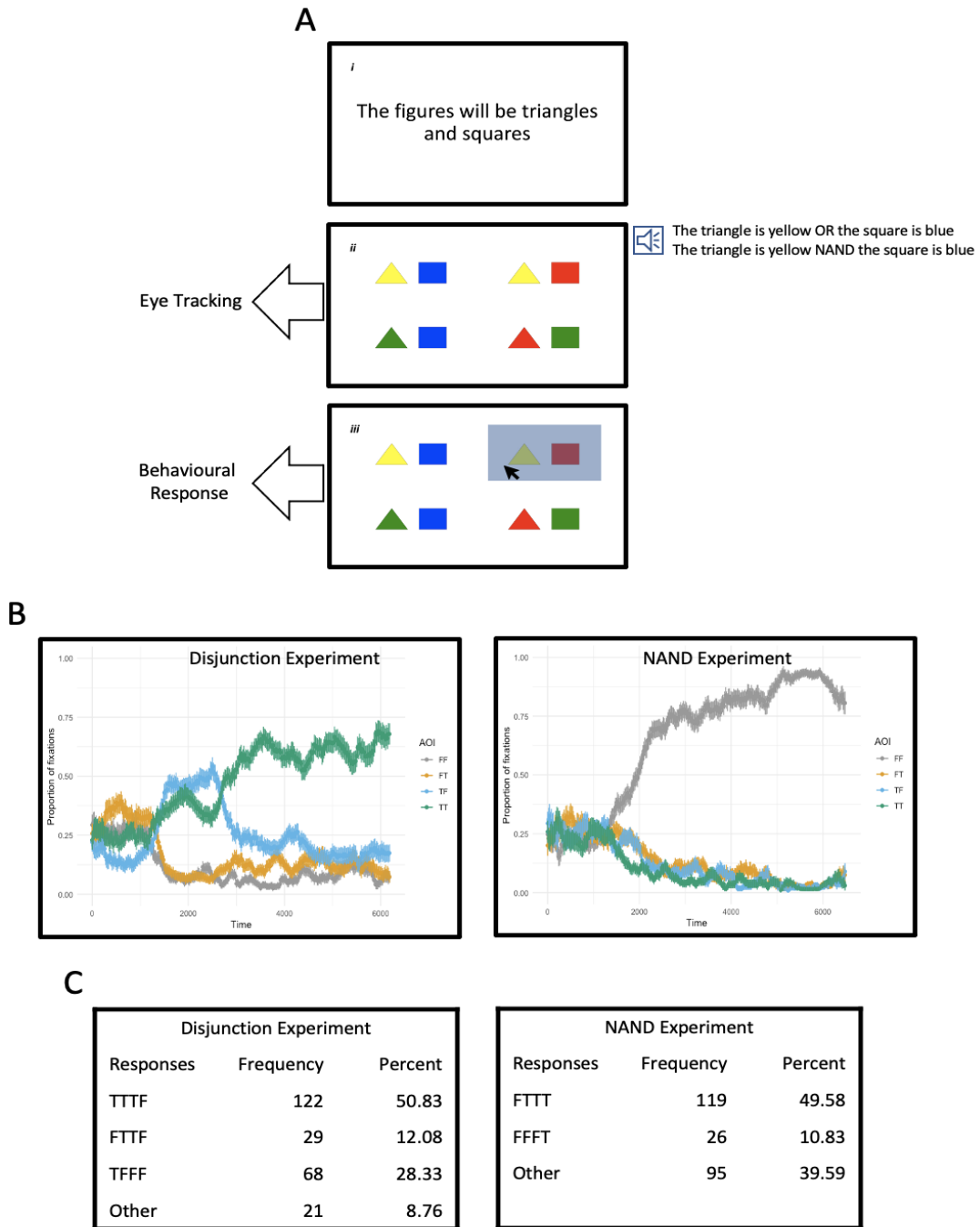| NAND Experiment		
Responses	Frequency	Percent
FTTT	119	49.58
FFFT	26	10.83
Other	95	39.59

Figure 1: Interpreting and thinking with existent and non-existent verbal connectives. **(A)** (i)

Panel specifies the figures participants will see in the trial. (ii) Each quadrant in the experimental panel presents two figures in diverse combinations of colours. An audio of a sentence describing one or more quadrants is played soon after the figures appear, as shown to the right of the panel. Each quadrant represents a combination of the truth values of two propositions, as in the two first columns of Table 1 (namely, the combinations TT, TF, FT, and FF), while the sentence played to participants is a linguistic representation of two propositions ("clauses", in language) mediated by a logical connective (a linguistic coordinator). The particular linguistic connective used in each experiment signals a specific interpretation —i.e., a unique truth table, as shown in columns 3–7 of Table 1 (e.g., TFFF is the truth table, or meaning, of the connective *conjunction*). In Experiment 1 participants would hear a disjunctive sentence such as *the triangle is yellow OR the square is blue*, while in Experiment 2 participants would first undertake a learning session to expose them to the meaning of a made-up word standing for the unlexicalised connective NAND and then in the experimental session they would hear sentences such as *the triangle is yellow NAND the square is blue* (see **Materials and Methods** for details). The tracking of eye movements starts from the beginning of the sentence and continues for a further 3 seconds after the sentence finishes. (iii) At the end of the eye-tracking, the mouse pointer is activated and participants are asked to select all the quadrants that match the sentence. **(B)** Proportion of fixations to each quadrant (TT, TF, FT, and FF) for a duration of circa 6500 milliseconds (3500 ms for the longest sentence plus 3000 ms of 'looking time'). In Experiment 1 (left panel), participants mostly fixated on the TT and TF quadrants after c. 1000 ms and on mostly TT only after 3000 ms. In Experiment 2 (right panel), participants fixated almost exclusively on the FF quadrant from c. 2000 ms. **(C)** Behavioural responses, where the TTTF pattern, for instance, specifies that participants had selected quadrants TT, TF, and FT (for reference, see Table 1). In Experiment 1 (table on the left), participants selected the inclusive interpretation of disjunction (TTTF) more often than the exclusive interpretation

21

(FTTF) or a 'conjunctive' reading (the TFFF pattern). In Experiment 2 (table on the right), participants selected the FTTT pattern, the truth table of NAND, as the best interpretation for the non-existent verbal connective, in preference to the FFFT pattern, the truth table of NOR, another unlexicalised connective.
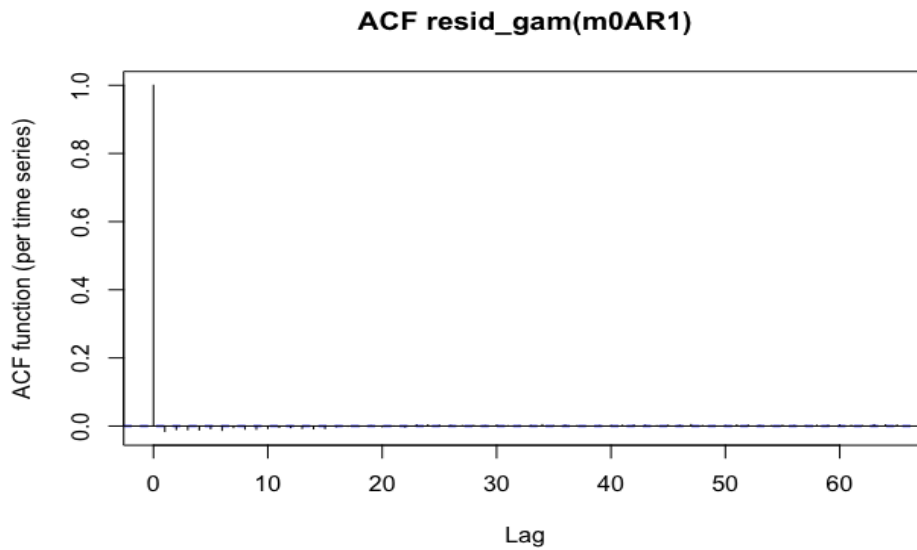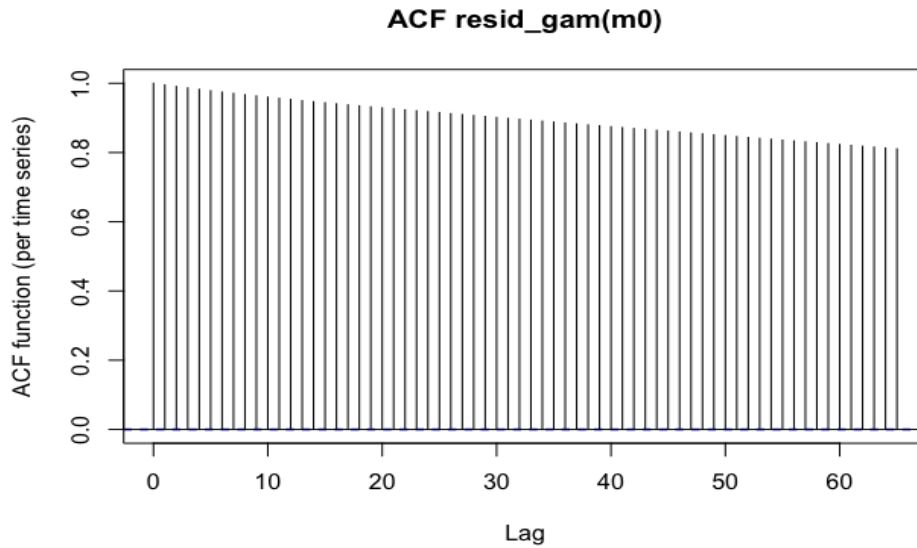
Figure 2: Amount of autocorrelation in two versions of model M0. Top panel shows the auto-correlation in a model without an AR-1 parameter (m0), while bottom panel shows the autocor-relation in a model with an AR-1 parameter (m0AR1). The autocorrelation is naturally '1' at time lag 0 (i.e., each point has a correlation of 1 with itself), decreasing therefrom. The height of the second line indicates the amount of autocorrelation at lag 1.
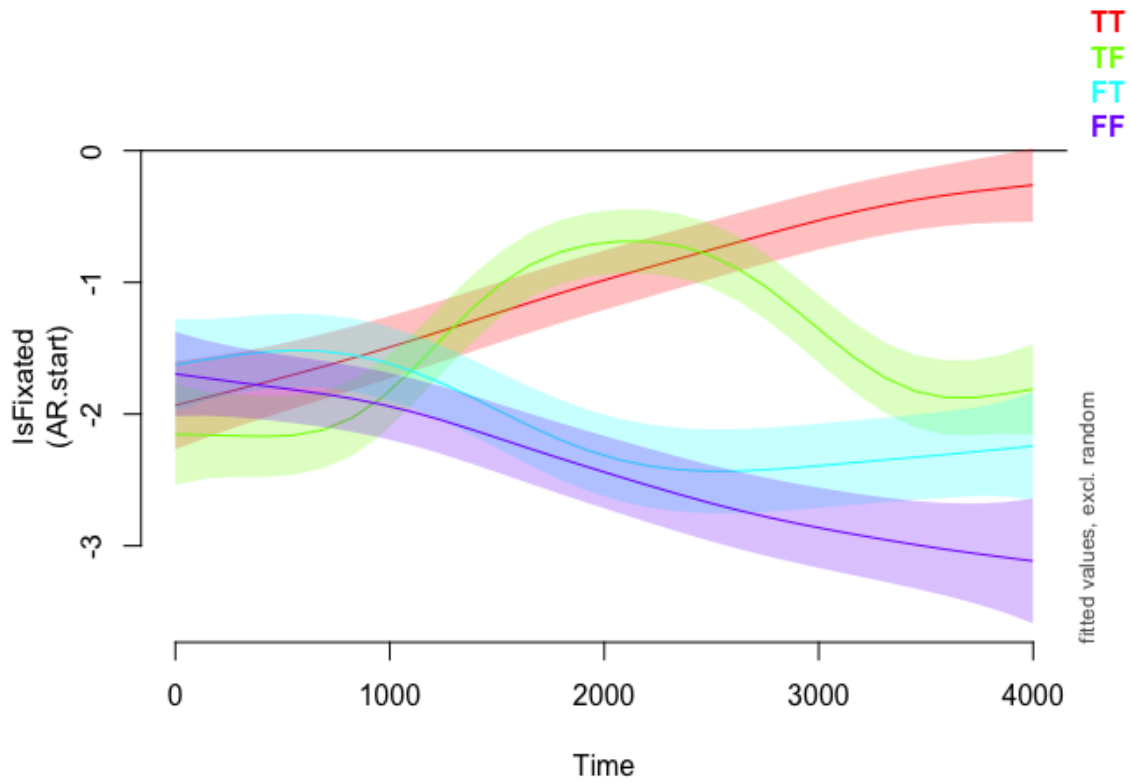
Figure 3: Non-linear smooths (fitted values) for fixations to each quadrant (TT, TF, FT, and FF) of model M2, Experiment 1, for a duration of 4000 milliseconds. The pointwise 95%-confidence intervals are shown by shaded bands.
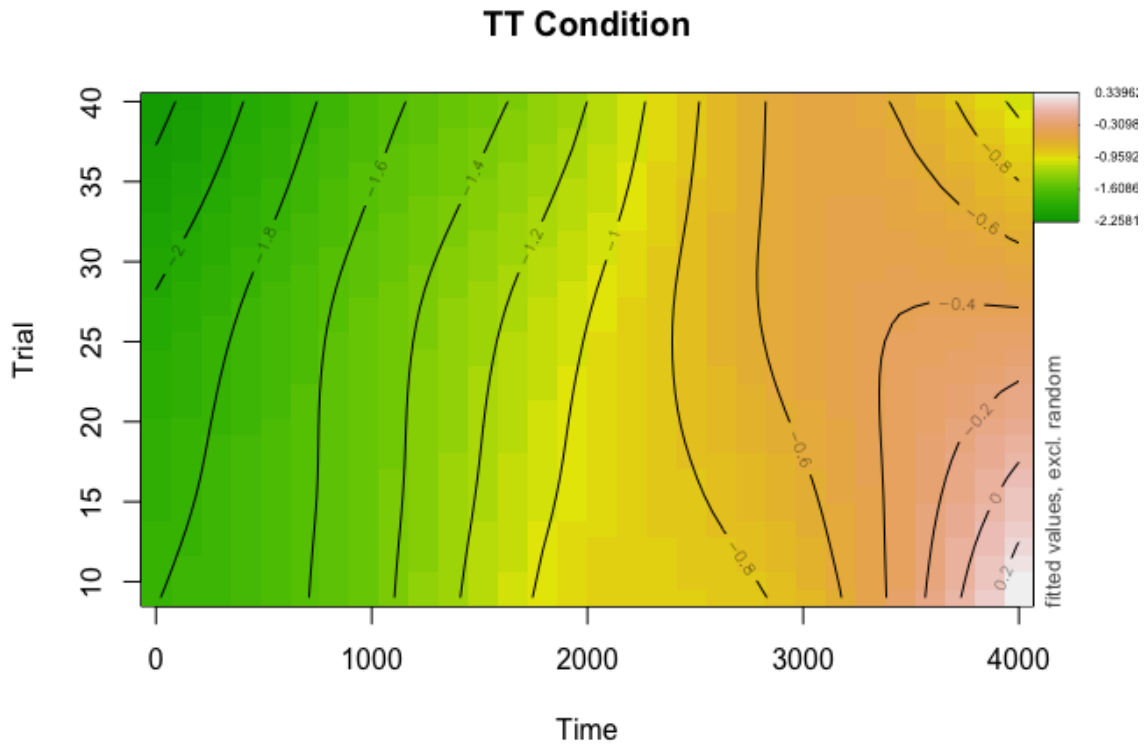
Figure 4: Contour surface for the Time x Trial interaction on the TT condition from model M2, Experiment 1. As the legend on the top-right corner states, areas in green indicate a smaller likelihood that the TT condition is fixated on, while a change of hue from green to yellow and then to orange indicates an increase in likelihood for fixations on the TT condition. Thus, there is an increase in the likelihood that the TT quadrant is fixated on across time, though towards the end of the time-series there is an actual decrease in the likelihood that the TT quadrant is fixated on as the experimental session progress —that is, there is a decrease in likelihood from trial number 20 onwards towards the end of each trial.
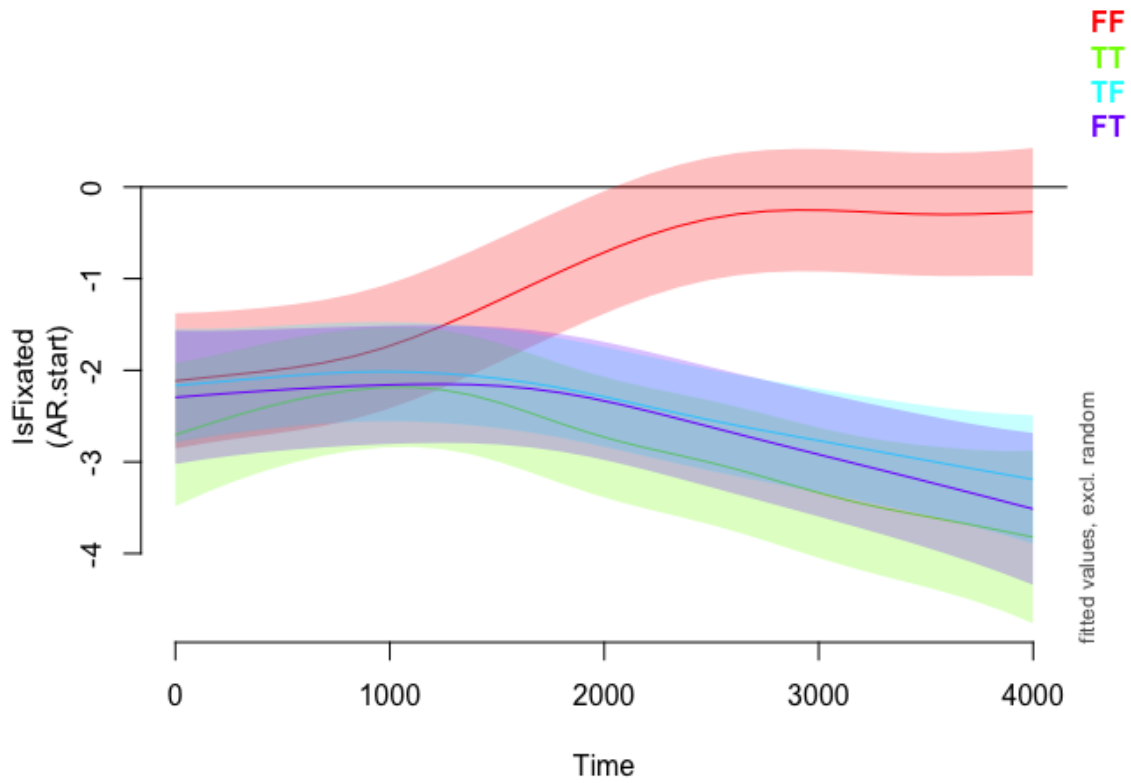
Figure 5: Non-linear smooths (fitted values) for fixations to each quadrant (TT, TF, FT, and FF) of model M1, Experiment 2, for a duration of 4000 milliseconds. The pointwise 95%-confidence intervals are shown by shaded bands.
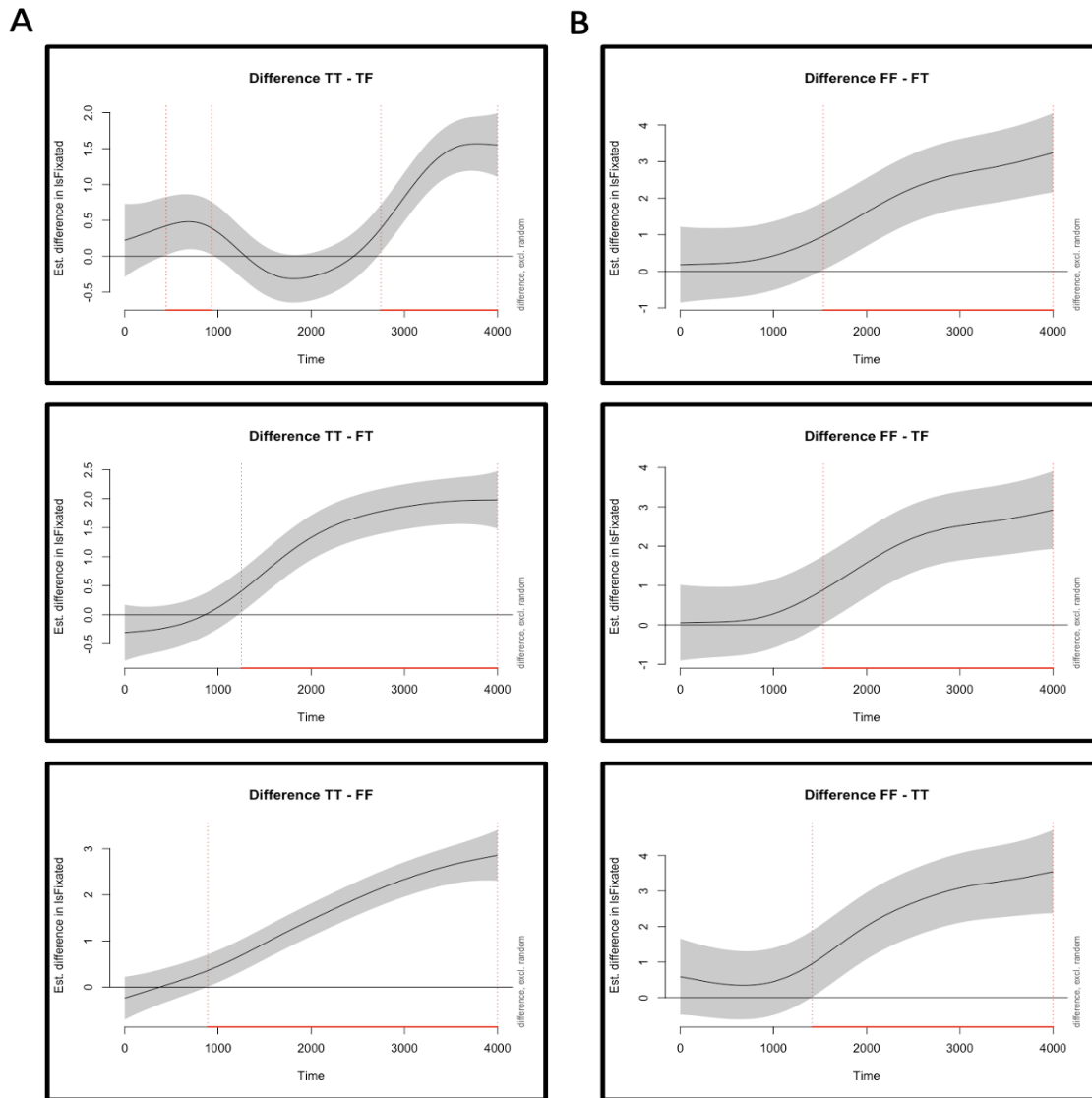
Figure 6: Difference curves, derived from the best-fit model of a generalised additive mixed-model analysis. The graphs show the comparison between the (non-linear) smooth of the quadrant with the most fixations (fixations are labelled as IsFixated on the y-axis) against each of the (non-linear) smooths of the other quadrants, with the gray solid line indicating the estimated difference. The shaded band represents the pointwise 95%-confidence interval; when the band doesn't overlap with the x-axis (i.e., the value is significantly different from zero), this is in-

27

dicated by a red solid line on the x-axis along with red vertical dotted lines. The graphs show fixations during the audio of the entire sentence in addition to an extra 640–1000 ms of "looking time" (see **Materials and Methods** for the selection of this time-window for the analyses). **(A)** Estimated differences from Experiment 1 between fixations to the TT quadrant and fixations to the TF, FT, and FF quadrants. The TT-TF comparison exhibits differences in two time windows, at window 445–930 ms, roughly around the time the first clause is being played in the audio, and at 2750–3999 ms, where the beginning of this time-window coincides with the end of the audio. The TT-FT contrast produces a difference at 1250–3999 ms, a window that (roughly) starts right after the connective has been presented, while the TT-FF comparison exhibits a difference at 890–3999 ms, where the beginning of this time-window precedes the presentation of the connective. **(B)** Estimated differences in Experiment 2 between fixations to the FF quadrant and fixations to the FT, TF, and TT quadrants. Comparisons show that all differences surfaced after around 1400 ms, when the connective has already appeared and the second clause is being presented.

3.2 Tables

Table 1: Logic's Truth Tables. Columns 1–2 show propositions P and Q and the different combinations of the truth values logicians assign to them in terms of whether the propositions describe a state of the world or not (T for true, F for false). The remaining columns specify the truth tables for the logical connectives Conjunction and NAND (columns 3–4), Inclusive and Exclusive Disjunctions (5–6), and NOR (7).

P	Q	$\wedge$	NAND	$\vee$	$\vee_e$	NOR
T	T	T	F	T	F	F
T	F	F	T	T	T	F
F	T	F	T	T	T	F
F	F	F	T	F	F	T

Table 2: Maximum Likelihood comparison between models M1 and M2, Experiment 1. M2 is preferred: lower fREML score (222582.690) and lower df (8.000).

Model	Score	Edf	Difference	Df
M1	-3587391	48	NA	NA
M2	-3809974	40	-222582.690	8.000

Table 3: Model Summary for M2, Experiment 1, reporting parametric coefficients (Part A) and estimated degrees of freedom (edf), reference degrees of freedom (Ref.df), F-values, and p-values for smooth terms (Part B).

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	-1.0590	0.0937	-11.3061	< .001
AOITF	-0.4270	0.1289	-3.3117	< .001
AOIFT	-1.0331	0.1329	-7.7732	< .001
AOIFF	-1.3615	0.1406	-9.6826	< .001
B. smooth terms	edf	Ref.df	F-value	p-value
s(Time):AOITT	2.5717	3.2734	77.3748	< .001
s(Time):AOITF	7.0725	8.2467	131.2745	< .001
s(Time):AOIFT	5.5059	6.8240	32.3071	< .001
s(Time):AOIFF	3.5746	4.6142	45.7139	< .001
s(Trial):AOITT	1.0163	1.0215	1.1134	.434
s(Trial):AOITF	1.0979	1.1349	0.0486	.822
s(Trial):AOIFT	1.2755	1.3917	2.3753	.129
s(Trial):AOIFF	1.0312	1.0468	1.1054	.317
ti(Time,Trial):AOITT	3.1402	3.6852	8.8481	.028
ti(Time,Trial):AOITF	1.0841	1.1605	0.0124	.981
ti(Time,Trial):AOIFT	1.4577	1.7883	3.7415	.217
ti(Time,Trial):AOIFF	1.0819	1.1612	0.5425	.525
s(Time,Event):AOITT	313.8023	1952.0000	1429.0621	< .001
s(Time,Event):AOITF	182.2498	1951.0000	2070.8300	< .0001
s(Time,Event):AOIFT	199.0825	1951.0000	951.4488	< .001
s(Time,Event):AOIFF	101.6644	1951.0000	98240.7144	< .001

Table 4: Maximum Likelihood comparison between models M1 and M2, Experiment 2. Chi-square test of fREML scores.

Model	Score	Edf	Difference	Df	p-value	Sig.
M2	-3800026	40	NA	NA	NA	NA
M1	4007992	48	207966.698	8.000	$< 2e - 16$	***

Table 5: Model Summary for M1, Experiment 2, reporting parametric coefficients (Part A) and estimated degrees of freedom (edf), reference degrees of freedom (Ref.df), F-values, and p-values for smooth terms (Part B).

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	-1.0634	0.3137	-3.3896	< .001
AOITT	-1.7535	0.4345	-4.0355	< .001
AOITF	-1.3671	0.4000	-3.4181	< .001
AOIFT	-1.3437	0.4137	-3.2476	.0012
B. smooth terms	edf	Ref.df	F-value	p-value
s(Time):AOIFF	4.8975	5.9350	59.5571	< .001
s(Time):AOITT	4.2796	5.3850	32.6855	< .001
s(Time):AOITF	2.9593	3.6781	12.9396	.009
s(Time):AOIFT	3.0391	3.7723	14.4254	.005
s(Trial):AOIFF	2.6887	3.3426	6.0813	.136
s(Trial):AOITT	1.2864	1.5064	1.7189	.212
s(Trial):AOITF	1.0488	1.0951	0.4639	.553
s(Trial):AOIFT	4.0588	4.9972	11.4329	.045
ti(Time,Trial):AOIFF	1.0466	1.0918	0.1404	.781
ti(Time,Trial):AOITT	2.4493	3.2051	3.2910	.402
ti(Time,Trial):AOITF	1.0954	1.1873	0.4488	.616
ti(Time,Trial):AOIFT	1.5503	1.9095	0.7982	.700
s(Time,Subject):AOIFF	43.2222	125.0000	181.1185	< .001
s(Time,Subject):AOITT	11.3449	125.0000	42.7670	< .001
s(Time,Subject):AOITF	11.0574	126.0000	48.2295	< .001
s(Time,Subject):AOIFT	19.5463	125.0000	68.6718	< .001
s(Time,Item):AOIFF	28.1990	143.0000	62.8804	< .001
s(Time,Item):AOITT	31.5330	143.0000	99.6503	< .001
s(Time,Item):AOITF	43.6394	144.0000	111.7409	< .001
s(Time,Item):AOIFT	40.7337	144.0000	100.4008	< .001

References

35. D. J. Barr, *Journal of Memory and Language* **59**, 457 (2008).

36. S. Donnelly, J. Verkuilen, *Journal of Memory and Language* **94**, 28 (2017).

37. D. Mirman, J. A. Dixon, J. S. Magnuson, *Journal of Memory and Language* **59**, 475 (2008).

38. R. H. Baayen, J. van Rij, C. de Cat, S. Wood, *Mixed-Effects Regression Models in Linguistics*, D. Speelman, K. Heylen, D. Geeraerts, eds. (Springer International Publishing AG, Cham, Switzerland, 2018), pp. 49–70.

39. S. N. Wood, *Generalized Additive Models: An Introduction with R* (Chapman and Hall/CRC, London, England, 2017), second edn.

40. R. H. Baayen, S. Vasishth, R. Kliegl, D. Bates, *Journal of Memory and Language* **94**, 206 (2017).

41. G. Montero-Melis, T. F. Jaeger, *Bilingualism: Language and Cognition* **23**, 602 (2019).

42. D. J. Lobina, J. Demestre, J. E. García-Albea, M. Guasch, *Linguistics and Philosophy* (2021).

43. V. Porretta, A.-J. Kyröläinen, J. van Rij, J. Järvikivi, VWPre: Tools for preprocessing visual world data (2016). Version 1.2.4, updated 2020-11-28.

44. J. van Rij, M. Wieling, R. H. Baayen, H. van Rijn, itsadug: Interpreting time series and autocorrelated data using gamms (2020). R package version 2.4.

45. M. Wieling, *Journal of Phonetics* **70**, 68 (2018).

46. J. van Rij, P. Hendriks, H. van Rijn, R. H. Baayen, S. N. Wood, *Pupillometry in Hearing Science* **23**, 1 (2019).

47. W. T. Fitch, A. Friederici, *Philosophical Transactions of the Royal Society B — Biological Sciences* **367**, 1933 (2012).

48. N. Akhtar, M. Tomasello, *Developmental psychology* **33**, 952 (1997).