Investigating Humor in EEG: Pun-Based Jokes Elicit Anterior N400 and Posterior P600 Effects

Philipp Keim, Cara Oster, Markus Werning
Department of Philosophy II, Ruhr University Bochum
Universitätsstraße 150, 44870 Bochum, Germany
philipp.keim@rub.de, cara.oster@rub.de, markus.werning@rub.de

Abstract

This study explores how readers process different types of pun-based jokes by analyzing their responses to various forms of linguistic ambiguity. Specifically, we examined puns rooted in homonymy, polysemy, and the contrast between idiomatic and literal interpretations of idiomatic expressions. Using EEG, we measured ERPs elicited by the ambiguous elements of these jokes and their punchline. These measurements enabled us to assess how distinct ambiguity types influence the comprehension of punchlines. Furthermore, we compared reader responses to puns against nonsensical sentences and straightforward control sentences. We hypothesize that the differences among joke types will manifest in the relative N400 amplitudes associated with punchline words, providing insights into the neural mechanisms underlying humor comprehension while having more control over joke setups compared to previous EEG studies in this field of research.

Keywords: Humor; Jokes; Puns; EEG; N400; P600

Introduction

Jokes pose a challenge to theories of language processing because, unlike straightforward sentences, they involve discovering a false epistemic commitment (Hurley, Dennett, & Adams, 2011) and require reinterpretation of their content when encountering the punchline. One way to lead joke comprehenders into a false epistemic commitment are linguistic ambiguities preceding the punchline. We refer to the position of this ambiguity as the pivot position of the joke, as it marks the turning point that allows for two divergent interpretations. To explore these processes, we designed an electroencephalography (EEG) experiment investigating how various forms of ambiguity in pun-based jokes impact event-related potentials (ERPs) associated with humor processing. Specifically, we are interested in homonymous (HOM), polysemous (POL), and idiomatic expressions (IDM), in order to compare three pun types based on these ambiguities. Our experimental design also allows ERP comparisons across four sentence types, which are joke (JOKE), joke control (jokeCTRL), nonsense (NONSENSE), and nonsense control (nonsenseCTRL) sentences, which will be further described below.

Each joke follows Raskin's (1985) ideas on script overlap and script opposition, which he considers necessary and sufficient conditions for a text to be a joke. At the textual level, the ambiguous part in our jokes enables two readings of the scenario up to the punchline. This ambiguity supports two distinct readings, each compatible with a different meaning of the ambiguous part. The pivot position marks where the joke enables these divergent interpretations. For linguistic ambiguities (homonyms and polysemes), the pivot is usually one word, while for idioms, it may be a phrase interpretable either idiomatically or literally. To understand the joke, the listener must reanalyze its setup — especially the ambiguous element — to pivot towards the correct interpretation, making the punchline follow from the setup. Here the punchline acts as a script-shift trigger that forces the reader to perform a script-switch and leads to reinterpretation. This reinterpretation of the ambiguous pivot leads to its less expected reading, given the setup of the joke.

Cognitively, understanding the joke requires comprehenders to initially commit to a particular interpretation of the setup based on the contextually induced meaning of an ambiguous part, which we refer to as its induced sense. When encountering the punchline, the reader realizes that they made a mistake in committing to this induced sense, where another reading of the ambiguous part would have been more appropriate to make the punchline follow from the joke setup. This recognition of having made a false epistemic commitment leads to a disconfirmation of expectations. Consequently, the discovery and debugging of the erroneous commitment is rewarded with the feeling of mirth, which Hurley et al. (2011) equates with the pleasurable experience of humor.

Method

Experimental Design

The experiment was conducted in German and had the form of a grammatical judgment task. A 2×2 design was used, with an ambiguous (Amb) or non-ambiguous (NonAmb) word in the context sentence at pivot position, which was followed by the target sentence containing either punchline or control as sentence final word at target position.

$$\underset{\textbf{pivot position}}{\{\text{Amb}, \text{NonAmb}\}} \times \underset{\textbf{target position}}{\{\text{Punchline}, \text{Control}\}}$$

This design manipulation created four experimental conditions as illustrated in Table 1.

In each trial, participants saw one to three context sentences followed by a target sentence. Both context and target sentences were split into roughly equal chunks displayed in the center of the screen, while the ambiguous/non-ambiguous word and punchline/control were presented in isolation and served as ERP trigger. Participants then judged the sentence's grammaticality by pressing *YES* or *NO*, with key assignments balanced across trials (see Figure 1 for an example with the exact presentation times). This was intended to ensure that participants engaged with the stimuli thoroughly and to prevent them from clicking through the task without proper pro-

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

Table 1: Four conditions of our 2×2 experiment with ambiguous (Amb) and non-ambiguous (NonAmb) words at pivot position (in ***italics***) and punchline and control at target position (in **bold** print)

	Punchline at target position	Control at target position
Amb at pivot position	JOKE In my profession, I like ***open*** people. I am a **surgeon.**	jokeCTRL In my profession, I like ***open*** people. I am a **social worker.**
NonAmb at pivot position	NONSENSE In my profession, I like ***approachable*** people. I am a **surgeon.**	nonsenseCTRL In my profession, I like ***approachable*** people. I am a **social worker.**

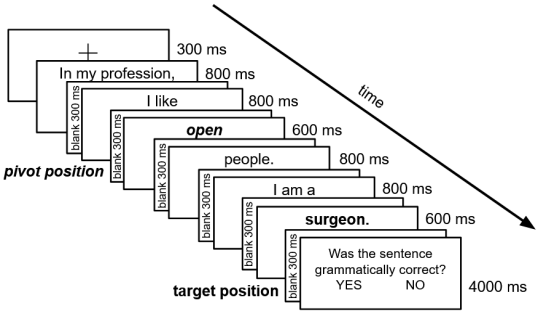

Figure 1: Surgeon Joke as Example for the Time Course of an Experimental Trial

cessing of the stimuli. The stimuli were presented in white on a black background.

Hypotheses

Coulson and Kutas (2001) state that the amplitude of the N400 correlates with the difficulty of lexical integration, indicating that greater integration challenges yield larger N400 amplitudes. Since processing jokes is believed to require more cognitive effort than straightforward sentences, we expect jokes to elicit a higher N400 amplitude. This aligns with findings from Chang, Ku, Wu, and Chen (2019), Coulson and Kutas (2001), and Xu, Nakanishi, and Coulson (2024), who reported a more negative N400 effect for jokes compared to control sentences. In our experiment, this is reflected in the comparison between JOKE and jokeCTRL. Therefore, we hypothesize that the N400 component at the target position will be more pronounced in JOKE than in jokeCTRL (H1). In our experiment jokeCTRL and nonsenseCTRL function as straightforward baselines to highlight the effects of incongruent endings. As can be seen in Table 1, this baseline comparison is possible, because JOKE and NONSENSE only differ in the final word compared to their straightforward counterparts.

Marinkovic et al. (2011) found that incongruent endings elicited higher N400 amplitudes than both congruent and funny endings, highlighting the N400 component's sensitivity to semantic incongruity. In our study, incongruent endings correspond to NONSENSE, where the punchline is incoherent within the preceding context, making it confusing or surprising. The congruent endings align with our nonsenseCTRL endings, which are straightforward and coherent, containing the non-ambiguous induced sense at pivot position. Lastly, funny punchlines correspond to our JOKE condition. Thus, when comparing NONSENSE and nonsenseCTRL, we expect a stronger N400 component for NONSENSE than for nonsenseCTRL (H2), and we predict that the amplitude of the N400 component will be larger for NONSENSE than for JOKE (H3). Besides an N400 component, Marinkovic et al. (2011) also found a P600 component in their study. They reported a more positive P600 component for funny punchlines compared to both congruent and incongruent ones, suggesting that the P600 reflects the effort to resolve the punchline's incongruity and establish coherence with the preceding context. In our experiment, this aligns with our JOKE condition, where resolving the punchline's incongruity is necessary to understand the joke, leading us to expect a pronounced P600 component here. Conversely, in jokeCTRL, which is straightforward and consistent with the context, there is no incongruity to resolve. Therefore, we hypothesize that the P600 component will be larger in JOKE than in jokeCTRL (H4). Similarly, Xu et al. (2024) found larger P600 effects for jokes compared to straight and expected sentences, further supporting the link between the P600 component and incongruity resolution in humor.

While developing hypotheses for comparisons between joke types, we found limited research directly addressing ambiguity types in jokes and instead relied on general theories about idioms, homonyms, and polysemes. Idioms, which are multi-word units characterized by being non-compositional, usually fixed, and involving figurative language (Wagner, 2021) are particularly relevant for pun-based jokes, as their figurative or idiomatic meanings often deviate considerably from their literal ones, providing an ideal ground for ambiguity and reinterpretation. In this regard, we reference Gibbs' (1980) direct access hypothesis for idiom processing, assuming that idiomatic readings are predominantly accessed before the literal readings. This preference makes deviations from the idiomatic reading more unexpected and likely more cognitively demanding. Based on Kutas, Van Petten, and Kluender (2006) and Hagoort (2007), who associate the N400 component with retrieval difficulty and contextual integration, we predict that processing idioms literally (rather than idiomatically) will yield higher N400 amplitudes. In our experiment, IDM puns are structured in a way that the reader is required to switch to the less expected literal reading to understand the

joke, unlike HOM and POL puns, where meanings are more equally accessible. Consider the following idiomatic pun: *My father always said he put blood, sweat, and tears into his work. He was a good man, but a terrible cook.* The punchline hinges on a reinterpretation of the figurative reading as literal, exemplifying the kind of cognitive shift our IDM puns require. Consequently, IDM puns, which involve a shift from the dominant idiomatic to the literal meaning, should demand more cognitive effort, leading to higher N400 amplitudes than HOM or POL puns (H5). Additionally, we expect the P600 in IDM jokes to be more pronounced compared to other joke types (HOM, POL), since in the case of idioms the reinterpretation is structurally more demanding (H6). To elaborate, a comprehender reinterpreting the idiom *to put blood, sweat, and tears into something* must shift from the figurative sense of intense effort to a literal image of someone physically adding bodily fluids to their work. We assume that this process involves more of a structural reinterpretation than, for example, switching from *open* meaning "physically open" to *open* meaning "approachable".

Another factor influencing the understanding of our pun types is the distinct processing of homonyms and polysemes. Polysemes have multiple, closely related meanings, while homonyms coincidentally share orthographic and phonological form but have distinct, unrelated meanings (Löbner, 2003; Maienborn, von Heusinger, & Portner, 2019). However, the line between them is often blurred, making a clear distinction difficult (Kroeger, 2018). For example, the polysemous surgeon joke from Table 1 relies on the related meanings of *open*, while a homonymous pun like *They live in a stable relationship, which is not uncommon for horses.* plays on the unrelated meanings of *stable*. Klepousniotou (2002) found that polysemes are processed faster than homonyms. This is likely because they share a common semantic core, which facilitates processing, while homonyms require choosing between meanings, which takes more time. Rodd, Gaskell, and Marslen-Wilson (2002) reported similar results, attributing slower processing of homonyms to competition between unrelated meanings. Buildings on these findings, we expect stronger N400 amplitudes for HOM jokes than POL jokes in our experiment, due to the greater cognitive effort needed to reinterpret unrelated meanings (H7).

So far, our hypotheses focused on ERPs at target position. However, since we assumed that the funniness of our jokes arises from the resolved ambiguity at pivot position after encountering the punchline, we also explored possible effects at pivot position. Drawing on the previously mentioned processing differences of homonymy, polysemy, and idioms, we also examine these ambiguities at the pivot position. We assume that the keyword of an idiom, which is the idioms final word and the word we measured on, will elicit a lower N400 amplitude relative to the single ambiguous word of a homonym or polyseme. This is because idioms, as multi-word units with fixed meanings, create strong expectations for the final word. For instance, upon reading "kick the" in "kick the bucket", the comprehender might anticipate "bucket" as expected continuation. Kuperberg and Jaeger (2016) assume that the more expected a word in a context with multiple possibilities is, the lower its N400 relative to less expected words, which is why we expect a lower N400 amplitude in IDM compared to HOM and POL (H8). This allows us to investigate how the features of the ambiguity types relate specifically to the pivot position. Hypotheses and expected results in short:

at target position:
- H1 N400(JOKE) > N400(jokeCTRL)
- H2 N400(NONSENSE) > N400(nonsenseCTRL)
- H3 N400(NONSENSE) > N400(JOKE)
- H4 P600(JOKE) > P600(jokeCTRL)

in JOKEs at target position:
- H5 N400(IDM) > N400(HOM) and N400(POL)
- H6 P600(IDM) > P600(HOM) and P600(POL)
- H7 N400(HOM) > N400(POL)

in JOKEs at pivot position:
- H8 N400(HOM) and N400(POL) > N400(IDM)

Materials

All jokes were of the type *pun*, meaning they are humorous wordplays (Attardo, 2018). We further grouped the jokes into three subcategories, i.e., according to their ambiguity type: homonymy (HOM), polysemy (POL), and idioms (IDM). Jokes varied in length, with up to four context sentences, 218 characters, and 35 words.

Alongside four experimental conditions, we added two filler conditions containing grammatical violations at either pivot or target position. All experimental trials were grammatically correct, allowing participants to focus on content rather than structure. This setup also helped to maintain attention by requiring *NO* responses to filler and *YES* responses to experimental trials during the grammatical judgment task.

Overall, 96 joke trials (JOKE) were generated, with corresponding counterparts for the three control conditions (jokeCTRL, NONSENSE, nonsenseCTRL). Stimuli were distributed across four lists, each containing 24 joke trials (8 per pun type: POL, HOM, IDM) and 24 trials per control conditions, totaling 96 experimental trials per list. Additionally, each list contained 48 pivot and 48 target fillers resulting in 192 trials per list, one half being filler and the other half experimental trials. Trials were distributed such that each experimental trial appeared only once per list but in different conditions (e.g., the surgeon joke was presented in JOKE in list 1, jokeCTRL in list 2, NONSENSE in list 3, and nonsenseCTRL in list 4). All participants attended four sessions, with a different list in each session, to avoid repetition of a trial in multiple conditions within the same session. This ensured that each participant saw all 96 joke trials and their corresponding control conditions across different sessions, thereby reducing memory effects. After four sessions, spaced at least two weeks apart, participants had seen all 96 experimental trials once and all 96 filler trials twice. Participants were randomly assigned to a list order.

Since the joke stimuli were self-constructed, we conducted an a priori online validation with a sample independent of

the EEG study. Participants rated joke and filler items on a 6-point funniness scale (1 = *not funny at all*, 6 = *very funny*) using a questionnaire. Grammatical errors in the filler items were corrected to allow for an adequate comparison. EEG participants completed the same questionnaire after their final session. In both groups, there was a significant difference between the funniness ratings of joke and filler items. Given these results, we can assume that our joke items were rated as being funnier than our filler items.

Participants

A total of 38 participants completed all four sessions of the experiment. Data of one participant was excluded because of their lack of attention during one session, i.e., less than 50 % of correct answers in the grammatical judgment task. Of the remaining 37 participants, 26 identified as female and 11 as male (age: 19 - 32, mean: 24.73, *SD*: 2.85). All participants were German native speakers who did not learn a second language as their mother tongue, had normal or corrected-to-normal vision, no neurological disorders, were right-handed and reported no use of medication during the study.

Procedure

At their first session, participants signed a written consent form, were screened for demographic criteria, and assessed for handedness using a translated version of the Edinburgh Handedness Inventory (Oldfield, 1971). Then, they completed two pretests, the Digit Span Test (DST) and the German computerized Reading Span Test (RST) (Van den Noort, Bosch, Haverkort, & Hugdahl, 2008), followed by the Autism Spectrum Quotient Questionnaire (AQ) (Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001).

The EEG measurement was conducted in an electrically and acoustically shielded cabin using NBS Presentation® Software for stimulus presentation. Participants sat in front of a shielded glass window with a computer screen behind it and a two-button response pad was placed in front of them. After setup of the electrode cap, they received written instructions and completed five practice trials. No feedback was given during the task. Participants were asked to read sentences carefully and respond to the grammatical judgment task, while minimizing movement. Although framed as main task, the grammatical judgment was designed to obscure the study objective. Each session included six blocks with breaks, totaling about 48 minutes net measurement time. The first session lasted about 3 hours due to questionnaires and pretests, session 2 to 4 were shorter (2 to 2.5 hours). At the end of session 4, participants completed a posttest questionnaire in which they rated the JOKE stimuli, using the same online questionnaire as in the a priori validation.

EEG Recording and Data Processing

EEG activity was recorded from 64 active electrodes (10-20 system) using a BrainAmp actiCAP system, with AFz as ground electrode and FCz as reference. Eye movements were tracked with four electro-oculogram electrodes (vertical: above/below right eye, horizontal: left/right temples). Impedance was kept below 5 kΩ. Data were sampled at 1000 Hz with a 10 s low cut-off filter and a hardware anti-aliasing filter. Processing was performed using Brain Vision Analyzer 2.2.0. An off-line band-pass filter 0.1 – 30 Hz (order 4) was applied, and data were down-sampled to 500 Hz. Trials were rejected if amplitudes exceeded 150 μV/150 ms, dropped below 0.5 μV/100 ms, or if the voltage step was increased above 50 μV/ms. Eye-movements were corrected by means of independent component analysis and data was re-referenced to the average of mastoid electrodes (TP9, TP10). Segments from 200 ms pre-target onset to 1000 ms post-onset were extracted and baseline correction used the 200 ms interval preceding stimulus onset. Segments with physical artifacts (amplitude $<$ -90 μV or $>$ 90 μV) were removed, and condition averages were generated for each participant.

Statistical Analysis

To assess significant differences in recorded ERPs across conditions, we performed a cluster-based permutation test (CBPT) using the Matlab FieldTrip package (Maris & Oostenveld, 2007; Oostenveld, Fries, Maris, & Schoffelen, 2011). This approach was chosen over a classical ANOVA due to its flexibility and fewer assumptions. While ANOVA requires predefined time windows and assumptions about effect timing and location, CBPT analyzes the entire epoch without segmentation, detecting both expected and unexpected effects while controlling for multiple comparisons without compromising statistical power (Groppe, Urbach, & Kutas, 2011). The analysis included all channels and focused on the epoch of 0 - 1000 ms post-onset. First, pairwise comparisons were conducted for words at target position across experimental conditions (JOKE, jokeCTRL, NONSENSE, and nonsenseCTRL). Furthermore, we compared the different pun types (HOM, POL, IDM) for JOKEs at target position. We also compared words at pivot position of individual joke types with other joke types, e.g. Amb_IDM vs. Amb_POL.

A tow-tailed dependent t-test compared participant averages (time $\times$ channel) between conditions. Significant data points ($\alpha = 0.025$) were grouped by time-spatial adjacency, and cluster-level statistics were computed by summing t-values. ERP averages were randomly exchanged and obtained cluster-level statistics were recalculated, using the maximum value as test statistics. This simulation was repeated for 10.000 permutations to evaluate the p-value (Maris & Oostenveld, 2007).

Results

When comparing sentence types, visual inspection of the grand averages suggests differences between all hypothesized comparisons on central-midline electrodes (Fz, Cz, Pz). As shown in Table 2, for JOKE vs. jokeCTRL the CBPT found a significant positive cluster from approximately 430 ms to the end of the epoch ($p = 0.0047$) and a significant negative cluster from around 450 ms - 750 ms post-onset ($p = 0.0055$).

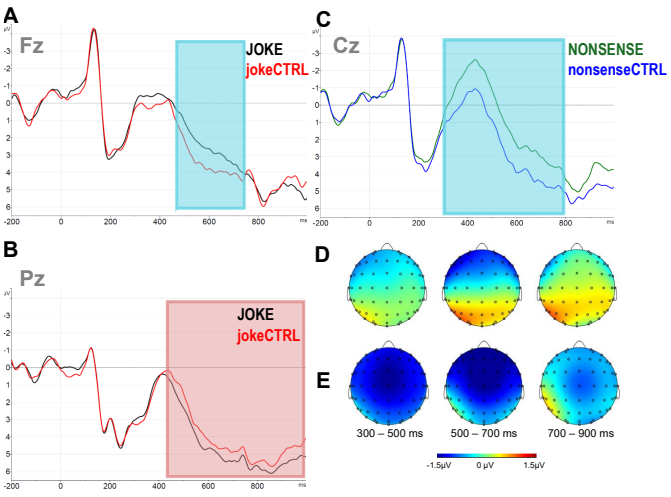

Figure 2: **Grand averages and topographic maps at target position**. **A** A comparison JOKE vs. jokeCTRL revealed a significant anterior N400 effect in a cluster from 450 - 750 ms (blue shading) in the CBPT. **B** Also, a significant posterior P600 effect from 430 - 1000 ms (red shading) was found. **C** A comparison NONSENSE vs. nonsenseCTRL revealed a significant central N400 effect from 300 - 800 ms (blue shading) **D** Topographic maps of JOKE vs. jokeCTRL. **E** Topographic maps of NONSENSE vs. nonsenseCTRL.

The positivity effect has a posterior distribution and is visible from about 500 - 1000 ms, while the negativity effect is distributed over anterior scalp sites and is most pronounced between 500 - 700 ms (see Figure 2). Regarding the comparison NONSENSE vs. nonsenseCTRL, we observed a significant negative cluster around the N400 time window (298 - 788 ms, $p < 0.001$). The negativity extends over the whole scalp centered at central-medial regions and decreases over time from posterior regions. When looking at NONSENSE vs. JOKE, the CBPT found a significant negative cluster between 160 - 1000 ms ($p < 0.001$). This long-lasting negativity extends over the whole scalp and is most pronounced at central-midline electrodes.

For comparisons between joke types, the CBPT found a significant negative cluster for HOM vs. POL from approximately 230 - 530 ms ($p = 0.0128$) with an anterior distribution. A significant positive cluster was found for IDM vs. HOM between around 200 - 350 ms ($p = 0.0339$), having an anterior distribution. For IDM vs. POL we found no significant differences (see Table 2 for a detailed overview of the results).

Apart from comparisons between sentence and joke types, we also compared sentences with regard to their pivot position (see Table 3). For Amb_IDM vs. Amb_POL, we observed a significant negative cluster between around 200 - 800 ms ($p < 0.001$) with central distribution. The comparison Amb_IDM vs. Amb_HOM also yielded a significant negative cluster with central distribution (216 - 638 ms, $p < 0.001$). No significant clusters were found in other comparisons.

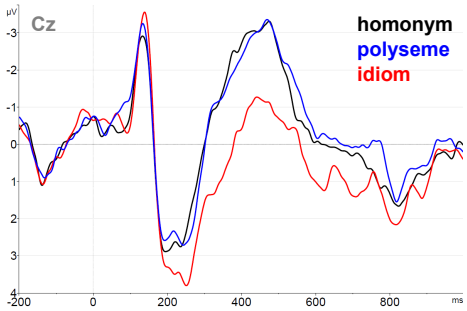

Figure 3: **Grand averages at pivot position for the three different ambiguity types, pooled JOKE and jokeCTRL.** In the N400 window idioms show a significantly lower negativity indicating the greater expectancy of the keywords in idioms.

Discussion

For the comparison between JOKE and jokeCTRL we predicted a more negative N400 component for jokes (H1). The significant negativity effect we found was visible between 450 - 750 ms but not in the time window typical for the N400. Additionally, this effect occurred at an anterior location, while the N400-effect is typically located at a centro-parietal location (Kutas & Federmeier, 2011). Therefore, we would interpret this effect as late left anterior negativity (LLAN) (Mayerhofer & Schacht, 2015). This effect is commonly found between 500 - 700 ms after stimulus onset and reflects additional processing efforts, which are required because of semantic re-interpretation. This re-interpretation is assumed to be the result of necessity to restore coherent discourse during joke processing (Mayerhofer & Schacht, 2015). Another, but not conflicting interpretation, is given by Coulson and Kutas (2001), who claim that this effect, which they refer to as sustained negativity over left lateral sites, indexes frame-shifting. For them, frame-shifting is necessary in order to get a joke and it can be interpreted in line with Raskin's theoretically claimed script-shifts (Raskin, 1985). However, it should still be noted that the left-lateralized distribution of the LLAN does not completely align with our finding, which also expands over the right anterior site. This could be the result of Coulson and Kutas (2001) distinguishing between good and poor joke comprehenders, a factor that our study did not address. We assume that our study included both types of comprehenders. Consequently, the outcomes reflect a mix of both groups, which may explain the difference in distribution compared to the findings of Coulson and Kutas (2001).

Additionally to the negativity effect, we also observed a significant positive cluster between JOKE relative to jokeCTRL (H4). This effect occurs in the time frame typical for the P600 component (426 - 1000 ms) and has the typical posterior distribution. This is in line with other studies (Marinkovic et al., 2011; Xu et al., 2024), which attribute this effect to incongruity resolution necessary in the JOKE condition relative to the congruous jokeCTRL condition.

In our comparison of NONSENSE vs. JOKE we predicted

Table 2: Significant clusters of pairwise comparisons in the CBPT for conditions at target position

Comparison	Polarity	Time (ms)	Distribution	Significance
JOKE vs. jokeCTRL	positive	426 - 1000	posterior	$p = 0.0047$
	negative	448 - 748	anterior	$p = 0.0055$
NONSENSE vs. nonsenseCTRL	negative	298 - 788	central	$p < 0.001$
NONSENSE vs. JOKE	negative	160 - 1000	central	$p < 0.001$
HOM vs. POL for JOKEs	negative	230 - 524	anterior	$p = 0.0128$
IDM vs. HOM for JOKEs	positive	196 - 358	anterior	$p = 0.0339$

Table 3: Significant clusters of pairwise comparisons in the CBPT for ambiguity types at pivot position

Comparison	Polarity	Time (ms)	Distribution	Significance
Amb_POL vs. Amb_IDM	negative	198 - 808	central	$p < 0.001$
Amb_HOM vs. Amb_IDM	negative	216 - 638	central	$p < 0.001$

an N400 effect and found a significant negative cluster ranging from approximately 150 ms to the end of the epoch (H3). Given the central scalp distribution and long duration of this effect, we interpret it as an N400 effect followed by a sustained negativity. This interpretation is supported by Mayerhofer and Schacht (2015), who also found extended N400 amplitudes when comparing jokes to nonsensical statements. This sustained negativity can be explained by the need to resolve an incongruity in both humorous and nonsensical statements. However, while incongruity resolution is possible in JOKE, it is not in NONSENSE. This successful incongruity resolution in JOKE is reflected in the posterior positivity effect observed between 430 - 1000 ms in the comparison with jokeCTRL. We interpret this as a P600 effect, as predicted (H4). In contrast, because the incongruity in NONSENSE cannot be resolved, it lacks this positive shift, making JOKE appear more negative in comparison. This also explains why no sustained negativity was found for NONSENSE vs. nonsenseCTRL, as nonsenseCTRL is coherent and does not require incongruity resolution. Thus, the significant negative cluster in NONSENSE vs. nonsenseCTRL can solely be explained due to NONSENSE being incongruent while nonsenseCTRL is not, confirming an N400 effect (H2). This aligns with Mayerhofer's and Schacht's (2015) findings.

For joke types, we predicted a higher N400 effect for idioms when comparing IDM vs. HOM, and IDM vs. POL on their respective target position (H5) as well as a P600 for the same comparisons (H6). We could not confirm H6 since the significant positive cluster between IDM vs. HOM is too short and too early (196 - 358 ms) to be any meaningful effect. Regarding H5, we argue that, contrary to the ideas we drew from the direct access hypothesis proposed by Gibbs, jokes relying on ambiguities caused by switching to the less expected literal meaning of an idiomatic phrase are no more surprising than those involving a shift from the contextually salient meaning of a homonymous or polysemous word to a less expected meaning. Consequently, we conclude that, at least in the context of joke processing, idioms do not have a more fixed way of representing contextually salient idiomatic meaning. The significant negative effect we found between HOM vs. POL had an anterior distribution but could be found in the typical time frame of the N400 effect. Therefore, we claim that this is an N400 effect with anterior distribution. We propose that this effect is the result of the greater semantic distance between the meanings of homonyms relative to the meanings of polysemes (H7). In order to switch from the induced sense in the homonymous pun to the sense necessary to get the joke, the comprehender has to perform a cognitively more demanding task compared to the switch necessary to get the polysemous pun. This relative difficulty is reflected in the anterior N400 effect in HOM vs. POL.

When comparing joke types at their respective pivot position, we found significant negative effects between comparisons involving idioms and other types of ambiguity. For Amb_IDM vs. Amb_POL and Amb_IDM vs. Amb_HOM, we believe the centrally distributed negativities are N400 effects (H8). Since the N400 amplitudes for idioms were significantly smaller than those for homonyms and polysemes, this hypothesis could be confirmed.

For future research on puns and humor in general, our experiment highlights that systematic differences in the setup of puns can influence the processing of their punchlines. When designing and interpreting experiments, it is important to account for the fact that the cognitive mechanisms involved in the script-shift required for joke comprehension may be influenced not only by different types of ambiguities but also by other structural elements of the joke. These factors may currently be understudied. To address this gap, future studies could explore how differences in the setup might affect the processing of punchlines across different types of jokes. For example, experiments sometimes distinguish between Theory of Mind (ToM) humor and semantic humor (Samson & Hegenloh, 2010; Manfredi et al., 2020). Similar setup-related effects, like those observed in our study, may also be present in ToM jokes, suggesting a potential area for further exploration.

Acknowledgments

We would like to thank our participants, our EEG-Lab team and the Bochum Chair of Philosophy of Language and Cognition for making this experiment possible.

References

Attardo, S. (2018). Universals in puns and humorous wordplay. In E. Winter-Froemel & V. Thaler (Eds.), *Cultures and Traditions of Wordplay and Wordplay Research* (pp. 89–110). Berlin, Boston: De Gruyter. doi: 10.1515/9783110586374-005

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Males and Females, Scientists and Mathematicians. *Journal of Autism and Developmental Disorders*, *31*, 5-17. doi: 10.1023/A:1005653411471

Chang, Y.-T., Ku, L.-C., Wu, C.-L., & Chen, H.-C. (2019). Event-related potential (ERP) evidence for the differential cognitive processing of semantic jokes and pun jokes. *Journal of Cognitive Psychology*, *31*(2), 131–144. doi: 10.1080/20445911.2019.1583241

Coulson, S., & Kutas, M. (2001). Getting it: human event-related brain response to jokes in good and poor comprehenders. *Neuroscience letters*, *316*(2), 71–74. doi: 10.1016/S0304-3940(01)02387-4

Gibbs, R. W. (1980). Spilling the beans on understanding and memory for idioms in conversation. *Memory & Cognition*, *8*(2), 149-156. doi: 10.3758/BF03213418

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology*, *48*(12), 1711-1725. doi: 10.1111/j.1469-8986.2011.01273.x

Hagoort, P. (2007). The memory, unification, and control (MUC) model of language. In A. S. Meyer, L. Wheeldon, & A. Krott (Eds.), *Automaticity and Control in Language Processing* (p. 243-270). Psychology Press.

Hurley, M. M., Dennett, D. C., & Adams, R. B. (2011). *Inside Jokes: Using Humor to Reverse-Engineer the Mind*. The MIT Press. doi: 10.7551/mitpress/9027.001.0001

Klepousniotou, E. (2002). The Processing of Lexical Ambiguity: Homonymy and Polysemy in the Mental Lexicon. *Brain and Language*, *81*(1), 205-223. doi: 10.1006/brln.2001.2518

Kroeger, P. (2018). *Analyzing meaning: An introduction to semantics and pragmatics*. Language Science Press. doi: 10.5281/zenodo.1164112

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32-59. doi: 10.1080/23273798.2015.1102299

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual review of psychology*, *62*, 621–647. doi: 10.1146/annurev.psych.093008.131123

Kutas, M., Van Petten, C. K., & Kluender, R. (2006). Chapter 17 - Psycholinguistics Electrified II (1994–2005). In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of Psycholinguistics* (Second Edition ed., p. 659-724). London: Academic Press. doi: 10.1016/B978-012369374-7/50018-3

Löbner, S. (2003). *Semantik: Eine Einführung*. Berlin, Boston: De Gruyter. doi: 10.1515/9783110904260

Maienborn, C., von Heusinger, K., & Portner, P. (Eds.). (2019). *Semantics - Lexical Structures and Adjectives*. Berlin, Boston: De Gruyter Mouton. doi: 10.1515/9783110626391

Manfredi, M., Proverbio, A. M., Sanchez Mello de Pinho, P., Ribeiro, B., Comfort, W. E., Murrins Marques, L., & Boggio, P. S. (2020). Electrophysiological indexes of tom and non-tom humor in healthy adults. *Experimental Brain Research*, *238*, 789–805.

Marinkovic, K., Baldwin, S., Courtney, M. G., Witzel, T., Dale, A. M., & Halgren, E. (2011). Right hemisphere has the last laugh: neural dynamics of joke appreciation. *Cognitive, Affective, & Behavioral Neuroscience*, *11*, 113–130. doi: 10.3758/s13415-010-0017-7

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. doi: 10.1016/j.jneumeth.2007.03.024

Mayerhofer, B., & Schacht, A. (2015). From incoherence to mirth: neuro-cognitive processing of garden-path jokes. *Frontiers in psychology*, *6*, 550. doi: 10.3389/fpsyg.2015.00550

Oldfield, R. (1971). The assessment and analysis of handedness: The edinburgh inventory. *Neuropsychologia*, *9*(1), 97-113. doi: 10.1016/0028-3932(71)90067-4

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience*, *2011*(1), 156869. doi: 10.1155/2011/156869

Raskin, V. (1985). *Semantic Mechanisms of Humor*. Dordrecht: Springer Netherlands. doi: 10.1007/978-94-009-6472-3

Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002). Making Sense of Semantic Ambiguity: Semantic Competition in Lexical Access. *Journal of Memory and Language*, *46*(2), 245-266. doi: 10.1006/jmla.2001.2810

Samson, A. C., & Hegenloh, M. (2010). Stimulus characteristics affect humor processing in individuals with asperger syndrome. *Journal of Autism and Developmental Disorders*, *40*, 438–447.

Van den Noort, M., Bosch, P., Haverkort, M., & Hugdahl, K. (2008). A Standard Computerized Version of the Reading Span Test in Different Languages. *European Journal of Psychological Assessment*, *24*(1), 35–42. doi:

10.1027/1015-5759.24.1.35

Wagner, W. (2021). *Idioms and Ambiguity in Context: Phrasal and Compositional Readings of Idiomatic Expressions*. Berlin, Boston: De Gruyter. doi: 10.1515/9783110685459

Xu, H., Nakanishi, M., & Coulson, S. (2024). Revisiting Joke Comprehension with Surprisal and Contextual Similarity: Implication from N400 and P600 Components. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).