11

Anaphora: Experimental Methods for Investigating Coreference

Elsi Kaiser

11.1 Introduction[1]

The topic of anaphora has played an important role in the development of syntactic theories. For example, reflexive binding is widely used as a diagnostic for c-command, the distribution of reflexives and pronouns has played a central role in theories of predicate-argument structure, and phi-feature (mis)matches between antecedents and anaphora inform investigations of feature checking. Binding Theory – which aims to characterize the distribution and interpretation of anaphora – is a central component of syntactic theorizing, and experimental syntacticians have investigated the predictions of Binding Theory in different syntactic configurations[2] and in various languages. At the same time, the processing of anaphora is of central interest for psycholinguistic work on the representations and mechanisms involved in the processing of anaphoric dependencies.

Methodologically, anaphora has been investigated using a variety of methods, with the choice of method often shaped by the empirical and

[1] In this chapter, I use the term "coreference" fairly broadly to mean situations when two linguistic expressions refer to the same entity and when one linguistic expression provides an antecedent for another linguistic expression. Under this usage, a reflexive and its antecedent are "coreferential," and a personal pronoun and its antecedent are also "coreferential," even if they cross a clause boundary. I do not intend the word "coreference" to be construed as excluding "binding" (see e.g. Bach & Partee (1980), Reinhart (1982, 1983a, 1983b) and others for an important semantic distinction that is often reflected in the terms *coreference* vs. *binding*). In this chapter, I intentionally use the term "coreference" in a broader way, but this terminological choice is not intended to detract from the existence of a semantic distinction between coreference and variable binding. See Frazier and Clifton (2000) and Carminati et al. (2002) for experimental work on this distinction. Most online psycholinguistic studies on anaphora have tended to focus on coreference relations, but see work by Frazier and colleagues, Kush, Lidz, and Phillips (2015) and Cunnings, Patterson, and Felser (2015) on variable binding.

[2] This chapter focuses mainly on reflexive and personal pronouns with sentence-internal antecedents – i.e. contexts where the distribution of pronouns and reflexives is (expected to be) governed by syntactic principles (e.g. Binding Theory). For work on cross-sentential (discourse-level) reference resolution, see e.g. Garnham (2001).

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

theoretical goals of a particular investigation. Acceptability judgments (also called grammaticality judgments; more on this distinction below) are one of the traditional tools of experimental syntax, and have been fruitfully used to investigate which antecedents are judged acceptable for a given anaphoric expression (e.g. to better understand the principles that constitute Binding Theory). However, as I discuss below, investigating coreference by means of acceptability judgments poses some very specific challenges that demand adjustments to the traditional acceptability judgment methods in order to avoid potentially uninterpretable results. I provide a detailed methodological discussion of how to make the necessary adjustments, caveats to watch out for, as well as related methods that can be used to complement acceptability judgment tasks.

Research using acceptability judgments and other *offline methods* (such as questions probing antecedent-choice) can be complemented by *online methods* such as self-paced reading, eye-tracking during reading and visual world eye-tracking. These methods allow researchers to tap into online processes that occur before comprehenders reach their final interpretation of the anaphoric expression, to detect fluctuations in processing load and potentially transient consideration (or lack thereof) of competing antecedents.

Ultimately, the choice of method depends largely on the research aims. Offline acceptability judgments can be used, for example, to assess whether a specific syntactically defined coreference configuration is judged acceptable by native speakers. But online methods are needed to assess questions such as whether processing an anaphoric expression only triggers activation of the Binding-Theory licensed antecedent, or whether syntactically inaccessible but featurally compatible referents are also activated. In this chapter, I mostly focus on acceptability-based methods, but I also include some comparative discussion on the merits of different methods such as eye-tracking.

Before continuing, let us acknowledge one of the hotly debated issues in experimental syntax in recent years, namely whether experimental methods are needed when investigating syntactic acceptability judgments. While some researchers argue that data from linguists' intuitions closely replicate data from psycholinguistic experiments, suggesting that experiments are not needed (e.g. Phillips & Lasnik 2003; Bornkessel-Schlesewsky & Schlesewsky 2007; Sprouse et al. 2013), others claim that psycholinguistic experiments either help or are necessary for obtaining reliable data (e.g. Gibson & Fedorenko 2013; Häussler & Juzek 2017) and that linguists' judgments differ from those of naïve native speakers (e.g. Dabrowska 2010). The present chapter will not address this debate and presupposes that readers are interested in an experimental approach to syntax in general and anaphora in particular (see Chapters 1 and 22 for further discussion).

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

Another broad, ongoing debate concerns the question of whether grammaticality itself is binary or continuous (and whether this is even a well-formed question). Researchers largely agree that, empirically, acceptability judgments are continuous (see e.g. Bader & Häussler 2010 for recent discussion): when participants are asked to judge the acceptability of sentences, their responses yield gradient data (see Chapter 3 for detailed discussion of gradience). However, as noted by Bader and Häussler (2010), "it is one thing to accept the gradience of grammaticality judgments but quite another thing to accept the notion of gradient grammaticality" (p. 276).

Before going further, it's worth noting that I follow Cowart (1997), Schütze and Sprouse (2013) and many others in talking about "acceptability judgments" rather than "grammaticality judgments." In generative grammar, the grammar is traditionally regarded as a mental construct that we cannot directly consciously access. When people report whether a given sentence is a possible sentence in their language (i.e. whether it is "acceptable"), their response is influenced not only by the grammar but also by factors such as real-world plausibility, frequency and ease of processing; see e.g. Schütze (1996) for discussion of these issues, which go back to the original competence–performance distinction. Furthermore, we cannot measure *directly* whether a given sentence is perceived to be acceptable by a participant: instead, we ask people to report their perception – more specifically, to report how acceptable they perceive that sentence to be – using various methods such as binary yes/no questions, multiple-point scales and so on (see also Bard et al. 1996; Schütze 1996). In this chapter, then, we are talking about acceptability judgments, known to be gradient (for further discussion of these issues, see Chapters 1–5).

The structure of this chapter is as follows. I first consider a fundamental issue that researchers conducting experimental work on anaphora need to address, especially if they are using acceptability judgment tasks: given that use of subscripts with naïve participants is typically avoided, how can researchers indicate the intended antecedent for an anaphor? If a researcher is interested in whether a particular coindexation relation is acceptable, how can this information be conveyed to naïve participants? Ignoring this important issue can yield uninterpretable data. I provide an in-depth discussion of different methodological approaches that can be used to indicate coreference when researchers want to elicit acceptability judgments, and compare these methods to other approaches that do not measure acceptability directly but rather ask participants to identify the preferred antecedents for anaphors.

The focus of this chapter is largely methodological. For an overview of some of the central theoretical issues regarding the processing of coreference (in particular reflexive pronouns), the reader is referred to Dillon (2014). For a typological overview of anaphora, see e.g. Huang (2000).

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

11.1.1 Using Acceptability Judgments to Investigate Coreference

Acceptability judgment tasks investigating syntactic issues outside the realm of anaphora typically involve participants rating the acceptability of individual sentences or the relative acceptability of sentence pairs. However, once we turn to the domain of anaphora, we need to distinguish two situations: does the researcher want to test (i) whether a given sentence is acceptable (assessing *sentence acceptability*) or (ii) whether a given anaphoric coindexation configuration is acceptable (assessing *coindexation acceptability*).

In many domains of experimental syntax, judgments of *sentence acceptability* provide the relevant data. However, in the domain of anaphora, the situation is slightly different: if a native speaker is asked about the acceptability of a sentence like (1a), she will presumably respond that (1a) is an acceptable sentence of English. Furthermore, when comprehending the sentence, she will presumably assign to the sentence the interpretation that *himself* is coreferential with Bob. However, the participant's response that "yes, this sentence is acceptable" does not provide a direct measure of what coreferential interpretation she gave to the reflexive pronoun. Thus, if a researcher wants to find out whether *himself* could also be coindexed with the matrix subject Alexander – a question that becomes especially relevant in languages with long-distance reflexives (e.g. Mandarin Chinese) – then simply asking a participant about the acceptability of the sentence as a whole fails to answer this question.

(1) a. Alexander said that Bob congratulated himself.

For a trained linguist, this situation has a straightforward answer: the use of subscripts (indices). In linguistic work, the intended coreferential interpretation – coindexation – is indicated by subscripted letters or numbers. Asterisks on the subscripts can be used to signal unacceptability of coindexation, as illustrated in (1b). Although theoretical linguists are familiar with these notational conventions, they may not be transparent for non-specialists. Thus, when conducting experiments with naïve, non-linguist participants, researchers typically opt not to use subscripts and signal the intended coreferential relation by other means (see Section 11.2).

(1) b. Alexander$_i$ said that Bob$_j$ congratulated himself$_{*i/j}$.

However, lack of any overt indication of coindexation patterns is not always a problem. For example, if a researcher is investigating reflexive pronouns in contexts where only one possible antecedent is present (and in a language where reflexives are known to not allow clause-external antecedents), then we can infer that if a participant judges the sentence to be acceptable, they also judge the one possible coindexing relation to be acceptable. This is the case for sentences like (2b): If a participant reports that this is an acceptable sentence of English, we can infer that they accept

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

the coindexation configuration where *herself* refers to Joan. In this case, assessing sentence acceptability allows us to assess coindexation acceptability.

(2) a. Joan respects her.
b. Joan respects herself.
c. She respects Joan.

But caution is called for, even in sentences that only contain one candidate antecedent, if the anaphoric form being tested is a personal pronoun – due to the possibility of a sentence-external antecedent. Consider (2a) and (2c), from Gordon and Hendrick (1997). Both of these sentences are unacceptable if the name and pronoun are coindexed, but fully acceptable if we assume that the pronoun is not coindexed with Joan – i.e. if a sentence-external referent is assumed for the pronoun.[3] Thus, simply asking participants about the acceptability of sentences like (2a, c) – without any coindexation information – will not allow a researcher to conclude that coreference between *Joan* and *her* in (2a) and *She* and *Joan* in (2c) is unacceptable. Thus, if our aim is to use acceptability judgments to investigate which coreferential relations are acceptable, we need to pin down what the intended coreferential relations are.

One solution is to explicitly tell participants to not consider any sentence-external referents. This was the tack taken by Lago et al. (2018) in recent work on possessive pronouns in German. Lago et al. used sentences with only one mentioned character and – to prevent participants from considering sentence-external referents – they explicitly instructed people to "provide their judgments only based on the current sentence context." In a speeded-acceptability task, the responses of native German speakers patterned as expected (i.e. they accepted sentences where the possessive pronoun gender-matched the mentioned referent more than sentences where the possessive pronoun did not gender-match the mentioned referent, 96% vs. 21%). The finding that participants largely reject sentences that would have been acceptable if a sentence-external referent had been considered indicates that the participants did indeed follow the instructions of only using/considering the sentence provided in the experiment.

[3] The challenges raised by the possibility of sentence-external antecedents can also occur with reflexives, see example (i). Pollard and Sag (1992) note that here, *John* can be coindexed with *himself*, despite being in a different sentence. Reflexives in certain syntactic environments (such as picture-NPs) are known to allow sentence-external referents. Thus, even with reflexives, absence of coindexation can result in data that is hard to interpret. The extent to which this complicates judgment tasks with reflexives depends on how widespread long-distance reflexives and "exempt" reflexives are in a given language. In English, we know from a large body of prior work that the distribution of so-called exempt reflexives is restricted to certain syntactic contexts.

(i) John$_i$ was furious. The picture of himself$_i$ in the museum had been mutilated. (from Pollard & Sag 1992)

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

11.2 Acceptability Judgments: How to Indicate Coreference without Subscripts

What about acceptability judgments of sentence containing more than one candidate antecedent – a configuration that is often a key test case for adjudicating between different theories (see Section 11.6)? In such a situation, if we cannot use subscripts to convey the intended coreference relation, how can participants be told what kind of referential dependency they are being asked to evaluate?

11.2.1 Enforcing Coreference by Means of Gender, Number or Person Cues

One means of spelling out the coreference relation that the researcher wants participants to assess is by using sentences where only one of the nouns featurally matches the anaphor. In this section, I discuss possible ways of enforcing coreference by means of (i) gender, (ii) number, or (iii) person cues – in essence, making sure that the anaphor only has one potential antecedent with matching phi-features.[4]

In English, where reflexives are marked for person, number, and gender, all of these features can be used to indicate the intended coindexation configuration. First, let's consider gender. (3a) allows us to test the acceptability of a local binder for the reflexive and (3b) allows us to test the acceptability of a non-local binder, by means of gendered names. In addition to names, nouns with definitional gender (*aunt*, *uncle*, *king*, *queen*), as in (3c, d), and professional roles with stereotypical gender (as determined by norming studies) have also been used (3e, f). There exists a lot of online processing work in English that has used gender features to investigate the processes involved in the real-time processing of reflexives.

Interestingly, processing work on English suggests that nouns with definitional gender and stereotypical gender pattern alike during real-time processing of anaphoric configurations but not in cataphoric configurations (e.g. Kreiner et al. 2008) – in essence, stereotypes can be overcome. Thus, for offline acceptability tasks, if one's aim is to pin down a certain interpretation by means of gender features, names and roles with definitional gender may provide clearer results than roles with stereotypical gender.[5]

[4] Another feature that could be used to constrain antecedent availability is animacy (specifically, use of inanimates), but creating minimal pairs of the type discussed below is often virtually impossible with inanimates.

[5] In some situations, however, researchers may want to use roles with stereotypical gender (as in (3e, f)) if they do not want to present participants with potentially ungrammatical sentences (compare (3d), which is ungrammatical, and (3f), which may be hard to process due to the stereotype violation but is not, strictly speaking, ungrammatical).

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

(3) a. *[names]* Alexandra said that *Bob* congratulated *himself/him.*
b. *Alexander* said that Barbara congratulated *himself/him.*
c. *[definitional gender]* The aunt said that the *uncle* congratulated *himself/him.*
d. The *uncle* said that the aunt congratulated *himself/him.*
e. *[stereotypical gender]* The beautician said that the *firefighter* congratulated *himself/him.*
f. The *firefighter* said that the beautician congratulated *himself/him.*

However, it is worth noting that, typologically speaking, reflexive pronouns with *gender marking* (like English *himself/herself*) are quite rare. Thus, for many languages, using *person or number cues* instead of gender may be helpful when designing sentences for acceptability studies involving coreference. This is demonstrated in (4a, b) for person cues (first-person "I" vs. a third-person name) and in (5a, b) for number cues.

(4) a. *[person cues]* I said that *Bob* congratulated *himself/him.*
b. *Bob* said that I congratulated *himself/him.*
c. *[number cues]* The children said that *Bob* congratulated *himself/him.*
d. *Bob* said that the children congratulated *himself/him.*

Thus, rather than subscripts, we are using person, number, and gender features to pin down the intended coreference relations between a reflexive and its antecedent. With this set-up, we can ask a participant to assess the acceptability of a particular coindexation relation.

It is important to acknowledge that the logic of this approach treats morphosemantic cues like gender as "hard-and-fast" cues that cannot be violated and thus provide an inviolable mean of pinning down the intended interpretation we want the participant to consider. However, one might wonder whether, during online processing, comprehenders ever "overlook" featural information and perhaps consider featurally mismatching antecedents to some extent.

So far, a considerable body of prior work seems to suggest that this is unlikely. There is some evidence that comprehenders may temporarily ignore structural constraints on anaphor resolution (Binding Theory) but there does not appear to be clear evidence that comprehenders would overlook phi-features such as gender, person, or number if those features are marked on the relevant anaphors in their language. (In fact, the "gender mismatch paradigm" hinges on the idea that presence of a gender-matching but structurally inaccessible referent improves the acceptability of sentences that otherwise *lack* a gender-matching antecedent – suggesting that comprehenders pay attention to gender cues. The picture is somewhat less clear for cross-clausal /cross-sentential pronouns; see e.g. Rigalleau, Caplan, and Baudiffier (2004) for a review; see also Fukumura, Hyönä, and Scholfield (2013).)

However, there also exists processing work – building on typological and theoretical work on the Feature Hierarchy – suggesting that *number features* are privileged over gender features during reference resolution (see Carminati 2005).

Thus, while gender, number, and person features can be useful in many languages for indicating the intended coreference relation, the question of how "hard-and-fast" these kinds of cues are in guiding comprehenders' reference resolution processes is not yet fully settled.

So far, we have considered how phi-features can be used to "pin down" the intended coreference relation for reflexives. What about personal pronouns? The approach of using gender, number, and person cues to preclude certain referents from consideration does not work as well with personal pronouns due to the possibility of a sentence-external referent. For example, under a sentence-external interpretation of *him*, all the examples in (4) are acceptable. Given that, for many theoretical issues, we want to be able to investigate and compare *both* pronouns and reflexives, this is a potential limitation of the "features-as-signalers-of-coreference" approaches. However, based on Lago et al. (2018), as discussed above, it may be possible to simply instruct participants to only consider referents mentioned in the provided sentences.

An example of an offline acceptability task using gender to disambiguate the referent comes from Foraker (2003), who investigated reflexives and pronouns in sentences like (5). Here, only *Megan* matches in gender with *her*/*herself*. Because the anaphoric form is part of a coordination, it is widely viewed as not being subject to Binding Theory – thereby raising the question of what are the interpretation preferences for pronouns and reflexives these configurations, and whether they differ from each other. Foraker manipulated the locality of the gender-matching antecedent (e.g. *Megan*) relative to the pronoun and reflexive:

(5) a. *Megan* wondered if Isaac had found out that Rick wanted to invite Sally and {*herself*/*her*} to the birthday party.
b. Isaac wondered if *Megan* had found out that Rick wanted to invite Sally and {*herself*/*her*} to the birthday party.
c. Rick wondered if Isaac had found out that *Megan* wanted to invite Sally and {*herself*/*her*} to the birthday party.

Participants rated the acceptability of sentences on a seven-point scale (1 = "nonsense/unacceptable," 7 = "makes perfect sense/fully acceptable"). Reflexives were rated more acceptable with local antecedents in the same finite clause (5c) than with antecedents in non-local positions (5a, b). Conversely, pronouns were rated more acceptable with non-local antecedents (5a, b) than local antecedents (5c). These results suggest that

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

reflexives and pronouns inside coordination are not as fully "exempt" from Binding Theory as is often assumed, since they still exhibit the standard locality effects even when inside coordination structures.

In closing, we need to acknowledge that, in some languages, disambiguation by means of person, number, or gender may not be easily achievable. For example, Chinese *ziji* 'self' does not mark gender, number, or person. In the next section, I discuss alternative ways of indicating the intended coreference relation when asking participants to assess sentence acceptability.

Before continuing, a word of warning: on a conceptual level, we need to keep in mind that participants are, in essence, often faced with two kinds of information when considering the kinds of sentences discussed in this section: (i) is the sentence acceptable under *some* syntactically licensed coindexation option and (ii) is the sentence acceptable under *the particular (experimenter-specified)* coindexation option? If the answer to (i) is "yes" but the answer to (ii) is "no" (e.g. consider example (1b)), one may wonder whether participants will be able to fully dissociate these two answers when rating the acceptability of the sentence.[6] Is there a risk of interference effects? In other words, is there a danger that participants will perceive a sentence as more acceptable due to the existence of a potential grammatically licensed antecedent (even if the one indicated by the experimenter is not actually an acceptable antecedent), and thus report artificially high acceptability scores? This appears to still be an open question.

11.2.2 Signaling Coreference by Typographic Means: Bold Font, Capitalization, Color, Boxed Font

The coreferential interpretation whose acceptability participants are asked to assess can also be indicated metalinguistically using means other than subscripts. A number of prior studies have used typographical means as a metalinguistic signal for coreference, rather than excluding certain coreferential options by means of gender, number, or person features.

In this section we consider four different typographic means of indicating the intended coreference relations: (i) bold font, (ii) all-caps, (iii) colored font, and (iv) use of a box around the anaphoric expression. These methods work well for materials presented in writing to literate participants, but (i) do not allow effects of prosody/intonation to be assessed,[7] (ii) are not suitable for investigating non-literate or pre-literate

[6] Note that studies investigating the Gender Mismatch Effect (GMME) typically use sentences where the answer to (i) is "no," i.e. where there exists *no grammatically licensed* antecedent that matches the anaphor in gender features (hence the term mismatch).

[7] Thus, they differ from the featural-disambiguation approach described in the preceding section, which allows for auditory presentation of the stimuli.

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

populations (e.g. young children), and (iii) may pose challenges for populations unfamiliar with thinking about language in metalinguistic terms.

The earliest work in this tradition, by Gordon and Hendrick (1997), used bold font to indicate the nouns whose coreferential relations were being investigated, as in (6). Among other things, Gordon and Hendrick set out to investigate how the core claims of Chomskyan Binding Theory correspond to the judgments of naïve participants.

(6) a. **Joan** respects **her.**
b. **Joan** respects **herself.**

In some of their studies, participants were instructed to simply indicate (in a binary manner) whether or not each sentence was acceptable if the two boldfaced elements corefer: Gordon and Hendrick state that "each sentence was accompanied by a check-off to indicate whether it would be acceptable if the boldfaced NPs it contained were coreferential" (Gordon & Hendrick 1997: 337). In addition, some studies used a six-point rating scale, labeled as follows: *1 Completely Unacceptable, 2 Unacceptable, 3 Just Barely Unacceptable, 4 Just Barely Acceptable, 5 Acceptable, 6 Completely Acceptable*. Participants used the scale to indicate the acceptability of the two bolded NPs being coreferential. (By labeling all points on the scale, Gordon and Hendrick depart from Cowart (1997: 71), who recommends only identifying the endpoints of the scale and leaving the points in the middle unlabeled. One of the reasons to only label the endpoints is that this can help yield interval (as opposed to ordinal) data; see Chapters 1 and 2 and Cowart (1997) for further discussion.)

A second typographic means of indicating coreference is capitalization. For example, Keller and Asudeh (2001) used a Magnitude Estimation task (see Section 11.3.3) to investigate the coreferential possibilities for pronouns and reflexives in a variety of syntactic positions, and used capitalization to indicate which elements should be interpreted as coreferential. Their instructions were: "Your task is to judge how acceptable each sentence is by assigning a number to it. By acceptability we mean the following: Every sentence will contain two expressions in ALL CAPITALS. A sentence is acceptable if these two expressions can refer to the same person." As a whole, Keller and Asudeh's work builds on Gordon and Hendrick (1997) but also investigated configurations traditionally viewed as being outside the purview of Binding Theory ("exempt from" BT), in particular picture-NP constructions (*picture of herself/her*). The coreference possibilities of pronouns and reflexives in picture-NPs have been the subject of much debate over the years, and by adopting an experimental approach, Keller and Asudeh show that while structural factors play a larger role than often assumed, the structural effects at play are not the ones that traditional Binding Theory leads us to expect.

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

In subsequent work, Kaiser, Nichols, and Wang (2018) used both bold font and underlining in a study investigating what kinds of anaphoric expressions can be used to refer to "imposters" such as *Mommy* and *Daddy* when used by parents to refer to themselves, as in (7). Imposters are expressions which are syntactically third person but semantically refer to the first-person speaker or the second-person addressee, e.g. *the present authors* when used by the authors to refer to themselves, *Mommy/Daddy* when used by parents to refer to themselves. We wanted to investigate the featural properties of pronouns (first- vs. third-person) that are coreferential with imposters and whether the acceptability of such coreference relations is influenced by the singular/plural distinction. We opted to use both underlining and bold font to ensure that the critical words would be clearly marked when displayed on different internet browsers.

(7) a. *Father says to child*: **Daddy and Mommy** have to finish {**their/our**} coffees.
b. *Father says to child*: **Daddy** has to finish {**his/my**} coffee.

A third typographic convention, namely colored font, was used by Temme and Verhoeven (2017) and by Moulton et al (2018), who color-coded the two elements whose availability for coreference was being tested. Temme and Verhoeven investigated cataphoric configurations like (8), where the pronoun precedes its (quantified) referent. They investigated German, with a focus on investigating what factors influence whether syntactic objects that are experiencers can bound by a preceding pronoun (8b, c). (I use bold font here for ease of exposition, but they used color to mark the relevant words.) They tested a variety of verb classes and case-marking configurations (e.g. accusative and dative shown in (8b, c)).

(8) a. **His** health worried **every patient**. (Reinhart 2002: 271)
b. Neulich haben die Meinungen **seiner** Schwester **jeden** verwundert.
'Recently the opinions of **his** sister astonished **everyone**.' (my translation)
c. Letztens haben die Träume **seiner** Kinder **jedem** gefallen.
'Lately **his** children's dreams pleased **everyone**.'

The question that Temme and Verhoeven (2017) posed to their participants was "Do you find the sentence acceptable under the condition that the highlighted words relate to the same person?" (their translation of the German original: "Finden Sie den Satz akzeptabel unter der Bedingung, dass sich die beiden markierten Wörter auf dieselbe Person beziehen?"). Participants were instructed to provide a binary "acceptable/not acceptable" response. The results show both syntactic and semantic considerations are relevant for cataphora. Both verb class (experiencer vs. agentive verbs) and case-marking (accusative vs. dative) modulate acceptability: backwards binding is more acceptable with datives and with experiencer verbs.

In related work, Moulton et al. (2018) presented the critical words in green in their investigation of cataphora in sentences like (9).

(9) Question: Who did John's wife hug? (object focus) OR Who hugged John? (subject focus)
Target sentence: **His** wife hugged **John** OR **His** wife hugged **him**.

(I again use bold font here for ease of exposition; they used color.) Moulton et al. were interested to see whether coreference between the possessive pronoun and the subsequent R-expression or pronoun was more acceptable when the R-expression/pronoun was focused or unfocused. The critical sentences were presented in writing (with two words in green) and were preceded by an auditorily-presented *wh*-question which focused the subject or object of the critical sentence. Participants gave a binary yes/no response to the question "Can the two parts in green refer to the same person?" The results suggest that being unfocused boosts the likelihood of coreference with the preceding possessive pronoun, which they link to QUD-related effects.

Finally, a fourth typographical option is to put a box around the anaphor, as done by Cunnings and Sturt (2014, 2018) in their investigation of pronouns and reflexives in co-argument positions and in non-co-argument positions (specifically, picture-NPs). It is important to point out that Cunnings and Sturt did not ask their participants to rate acceptability and thus their task differs from the others reviewed in this section: Their participants saw sentences like (10), and "were instructed to choose who they thought the boxed pronoun most likely referred to, and were given the options to choose person (A), person (B) or either of them" (Cunnings & Sturt 2018: 1246). This task asks participants to select an antecedent for the pronoun, an approach I discuss more in Section 11.4.

(10) a. The surgeon remembered that Jonathan had noticed $\boxed{\text{him}}$ near the back of the lunch queue.
b. The surgeon remembered about Jonathan's picture of $\boxed{\text{him}}$ near the back of the lunch queue.

This method could easily be adapted to fit an acceptability judgment task, by putting a box around the anaphor and another box around the antecedent being tested – thus paralleling the other typographical, metalinguistic approaches described in this section. This approach can also avoid concerns regarding color blindness or differences in how colors are displayed on different screens.

In sum, a variety of typographical means have been used to circumvent the need for subscripts while still indicating that two elements are to be construed as referring to the same person. However, like subscripts, these typographical means are inherently *metalinguistic*, and participants (a) need

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

to be told that their task is to assess whether the two typographically marked elements can “refer” to the same person and (b) presumably may also need some examples or additional explanation about what it means to “refer.” Thus, the instructions for this kind of task are not as intuitive as for some other methods.

A possible concern with some of the typographical methods is the risk of participants perceiving the marked words as being emphasized in some way. This can introduce a confound: if typographic emphasis renders the referent more salient/prominent/accessible – or indicates that it is contrastively emphasized – this could (i) render the referent more available as an antecedent for a reflexive or pronoun than might otherwise be the case or (ii) evoke other potential antecedents as part of the alternative set evoked by contrastive focus. Indeed, recent work by Fraundorf et al. (2013) and Maia and Morris (2019) suggests that words presented in all-capitals evoke alternatives, indicating that they are interpreted as being contrastively focused. These studies, however, did not look at reflexives, and it is unclear whether these kinds of contrast effects have arisen in the prior anaphor-resolution studies. Nevertheless, the risk of “boosting” or otherwise affecting the representation of the typographically marked referent is a potential complication associated with those typographical formats that are conventionally used to indicate emphasis in text (such as all caps, italics, bold font, and underlining). However, this is probably less of a concern with color coding and boxes, as those are not conventionally used for marking emphasis or contrast in text.

11.2.3 Signaling Coreference by Other Means: Linguistic and Visual Context

Researchers have also used non-typographic means to indicate the intended coreference relation that they want participants to evaluate. In this section I review two options: (i) providing additional linguistic context and (ii) providing a visual context. An example of specifying the intended coreference relation by means of linguistic context comes from Featherston’s (2002) investigation of German object-position pronouns, reflexives, as well as reflexives modified by the intensifier *selbst* (‘self’). He used a preceding context as well as an explicit paraphrase at the end of the sentence to clarify the intended meaning. This is exemplified in (11), where both the story context and the addition of the “i.e.” clause (in German *d.h.*, short for ‘das heisst’) indicate that Martin saw Martin, and not someone else. Participants were instructed to judge whether each sentence sounded natural, and indicated their naturalness ratings using the Magnitude Estimation method, discussed in Section 11.3.3.

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

(11) Martins neuer Bundeswehrhaarschnitt gibt ihm den Anschein eines Sträflings. Manche finden es jedoch gemein von mir, dass ich Martin sich im Spiegel gezeigt habe. (*d.h. Martin sah Martin)*
'Martin's new army haircut made him look like a convict. But some people thought it was mean of me that I showed Martin himself in the mirror. (*i.e. Martin saw Martin*).'

Another non-typographic approach uses the picture-verification task: Participants are presented with a pictorial depiction of the intended anaphoric relation, and indicate whether (or not) the sentence and the picture match (see also Chapter 13). For example, Kaiser et al. (2009) investigated the interpretation of reflexives and pronouns in picture-NPs with and without possessors (e.g. *picture of him/himself*, *Mary's picture of her/herself*). One of the questions we tested was which anaphoric relations are acceptable for pronouns and reflexives in sentences like (12), which have two potential antecedent candidates.

(12) John {told/heard from} Peter about the picture of {him/himself}.

We asked whether pronouns can refer to subject antecedents and reflexives to non-subject antecedents (contrary to what is typically assumed) and whether this is modulated by the discourse/semantic properties of that referent (source or perceiver of information), modulated by the verb. In a picture-verification task, participants were shown images with, for example, John and Peter and a framed picture of either John (subject antecedent) *or* Peter (object antecedent). Participants indicated (yes/no) whether the sentence matches the image. If participants respond that yes, the sentence matches the picture, this indicates that the anaphoric relation represented in the picture (subject or object antecedent) is acceptable for the pronoun or reflexive in the sentence. Our results show that a purely syntactic account is not sufficient: while pronouns elicit more "yes" responses for object antecedents and reflexives for subject antecedents, the proportion of "yes" responses for both is also modulated by referent's thematic role: reflexives exhibit a preference for sources of information, while pronouns prefer perceivers of information.

This method does not ask people to categorize sentences or anaphoric dependencies as "acceptable" or "unacceptable." Providing instructions to participants is thus relatively straightforward, as the metalinguistic notion of acceptability does not need to be mentioned. This method does not provide a *direct* measure of how unacceptable a certain anaphoric dependency is, but the proportion of "yes, sentence matches image" vs. "no, sentence does not match image" provides a measure of how willing participants are to accept different anaphoric dependencies.

Given its intuitiveness, it is not surprising that the picture-verification task originated in language acquisition research. For example, in their influential work on the acquisition of Binding Theory, Chien and Wexler

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

(1990) presented children with sentences like "Is Mama Bear touching her/herself?" and showed them images with Mama Bear and Goldilocks where Mama Bear was touching herself (subject antecedent) or was touching Goldilocks (sentence-external antecedent). The task was to answer "yes" or "no." This picture-verification method has been widely used in acquisition work, and has the advantage of not requiring metalinguistic explanation or reasoning about acceptability or notions like "refer."

11.3 How Do Participants Give Their Responses to Acceptability Tasks?

The preceding section reviewed different means of indicating the intended coreference relation to naïve participants, without using subscripts. In this section, we consider the nature of the dependent variable – how do participants indicate whether they judge the sentence to be acceptable? In addition to allowing participants to make distinctions at different grain sizes, the nature of the dependent variable (e.g. binary responses, *n*-point scales, continuous scales) also has important implications for the suitability of different statistical analyses (see e.g. Cowart 1997; Bard et al. 1996). Research on acceptability in general has used a wider range of dependent variables than research specifically on coreference. This section assumes that the broader methodological points are also relevant for coreference, although this is ultimately an empirical question.

11.3.1 Binary Responses

The simplest acceptability judgment tasks ask participants to provide a binary response, in line with the traditional categorical distinction of grammatical vs. ungrammatical. For example, Gordon and Hendrick (1997) asked participants to indicate by checking a box whether a sentence "would be acceptable if the boldfaced NPs it contained were coreferential" (Gordon & Hendrick 1997: 337; see also Section 11.2). A binary yes/no response can also be elicited by means of a picture-verification task, as described above.

A variant of the yes/no binary acceptability task is a speeded acceptability judgment task: participants see sentences word-by-word, with each word displayed for approx. 300–400 ms, and then provide a binary yes/no acceptability as fast as possible, with fast responses enforced by a response deadline of approx. 2 seconds (e.g. Bader & Häussler 2010; Wagers, Lau, & Phillips 2009). Lago et al. (2018) used speeded acceptability judgments to test processing of possessive pronouns in German by native speakers of English and Spanish. Lago et al.'s participants showed similar patterns in their response times to the speeded acceptability task and in a self-paced reading study, though the proportion of acceptances

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

("yes" responses) in the speeded acceptability task yielded somewhat different results. Thus, the reaction time component of speeded acceptability judgment tasks provides a sensitive means of tapping into processing.

11.3.2 Scales

Intuitively, people often feel that a binary yes/no response is not sufficient to express finer nuances of acceptability (which relates to the still-debated question of whether grammaticality is underlyingly continuous or inherently categorical, with variance attributed to performance factors; see Section 11.1). One alternative is to use *n*-point scales, where participants have the choice between more options than simply "acceptable" and "unacceptable." Empirically, researchers are still debating how the data obtained by means of binary responses compares to data obtained from *n*-point scales (see Chapter 2, as well as Weskott & Fanselow (2011), Sprouse & Almeida (2017), and Langsford et al. (2018), for different perspectives).

When using scales, one challenge to keep in mind has to do with how to interpret responses at the middle of the scale. Consider the five-point scale in (13). If a participant is *uncertain* about whether *him* and *John* can corefer, they would presumably choose 3. However, if a participant is very *certain* that this coreference relation has an *intermediate acceptability status*, they would presumably also choose 3. Thus, there are concerns about conflating certainty with acceptability (see also Ionin & Zyzyk 2014).

While it has been suggested that this challenge could be mitigated by using a scale with an even number of points (e.g. 1 to 6, as also shown in (13)),[8] it is not clear that an even-point scale fully solves the problem: Participants who are uncertain could still chose a number near the middle of the scale (3–4) and participants who are certain about a middling level of acceptability could also still choose a number near the middle of the scale. ((13) uses the metalinguistic labels "Unacceptable" and "Acceptable," which would need to be explained to participants beforehand.) Arguably, though, a clear midpoint (on an odd-numbered scale) is probably more likely to be construed as a means of responding "I don't know" (low certainty) than the middle of the scale on an even-numbered scale.

(13) **John** heard from Peter about the picture of **him**.
Unacceptable 1 2 3 4 5 Acceptable (odd number of points)
Unacceptable 1 2 3 4 5 6 Acceptable (even number of points)

[8] Gerken and Bever (1986)'s early work on pronouns used a four-point acceptability scale. Gordon and Hendrick (1997) used a six-point rating scale (see Section 11.2). It is worth noting that neither four- nor six-point scales have a midpoint, and thus participants are forced to make a decision about whether a sentence is more acceptable or unacceptable. (Five- and seven-point scales – also commonly used in linguistic experiments – offer a midpoint.)

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

As an alternative (or a supplement) to using an even-point scale, the experimenter could provide a separate "I don't know" answer choice that is distinct from the scale, or include an additional scale that asks participants to indicate their certainty/confidence in their answer (see Chapter 3, and also Rebuschat (2013), Ionin & Zyzyk (2014), Montrul, Dias & Santos (2011)). It is worth keeping in mind that use of a binary yes/no (acceptable/unacceptable) response avoids this scale midpoint complication. Thus, each method has its pros and cons.

In addition to considering scales anchored between acceptable and unacceptable, rating scales used to investigate coreference could also be constructed to be between two antecedent choices, as in (14). Such scales do not provide information about acceptability but rather about which referent is preferred as the antecedent (see e.g. Kaiser 2015).

(14) Lisa heard from Kate about the picture of herself on the wall.
Who is shown in the picture?
Lisa 1 2 3 4 5 6 Kate

Let us briefly consider another manipulation done by Gordon and Hendrick (1997) – namely the use of two kinds of instructions: what they term "reflective instruction" (which encouraged participants to reflect on the sentences before responding) and "immediate instructions" (which encouraged participants to respond immediately without reflecting). Although they found no main effects of instruction type, a closer look at the different conditions they tested suggests that in some conditions, the reflective instructions elicited stronger effects of c-command than the immediate instructions. (See also Cowart (1997: 57) for findings showing no effect of instruction type.)

11.3.3 Magnitude Estimation

In the 1990s a new method gained popularity which essentially offers participants an unlimited number of response options, namely Magnitude Estimation (ME) – a method that is widely used in psychophysics research (see Stevens 1975), and was introduced to linguistics by Bard et al. (1996) (see Chapter 2 for discussion). In this method, participants evaluate stimuli relative to a "reference stimulus," often a sentence of intermediate acceptability. Participants are instructed to judge the acceptability of the experimental stimuli relative to the acceptability of the reference stimulus. Crucially, participants can make as many distinctions as they perceive to be necessary.[9]

Keller (2000) and Keller and Asudeh (2001) were the first to systematically use Magnitude Estimation to probe coreference judgments.

[9] A related method is the thermometer task pioneered by Featherston. This method, however, differs in some important ways from Magnitude Estimation (see e.g. Featherston 2008 for more discussion).

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

Experiment 1 in Keller and Asudeh (2001) tested standard Binding Theory configurations using Magnitude Estimation methodology; Experiment 2 turned to a more contentious domain, namely picture-NPs (e.g. *the picture of her/herself*). Keller and Asudeh indicated coreference by means of capitalization. Their reference sentence was "Jill told the people HE trusts all about SAM" (see Keller 2000).

Experiment 1 in Keller and Asudeh (2001) replicates Experiment 3 of Gordon and Hendrick (1997), and shows that with uncontroversial Binding configurations (where Chomskyan Binding Theory makes clear claims about what coreference relations are grammatical for pronouns and reflexives), binary responses and Magnitude Estimation yield comparable acceptability rating data. Keller and Asudeh also investigated picture-NPs, and found that Magnitude Estimation can yield fine-grained information about coreference in this construction as well.

However, given the relative complexity of the method, one may wonder whether, in the words of Fukuda et al. (2012), this method is "worth the trouble." Indeed, a growing body of work suggests that Magnitude Estimation may not be "worth the trouble," in the sense of not yielding data that is more informative or stable than data obtained by means of a seven-point or a five-point scale (e.g. Bader & Häussler 2010; Weskott & Fanselow 2011; see also Langsford et al. (2018), who also tested Thurstonian methods from psychophysics; Thurstone 1927; Roberts et al. 1999; Fabrigar & Paik 2007). Although these methodological comparisons did not investigate coreference judgments, it seems reasonable to assume that their conclusions would also extend to the reference resolution domain.

11.4 Offline Methods Used in Work on Anaphors that Do Not Measure Acceptability

So far, we have focused on situations where the aim is to test whether a particular coreferential configuration – specified by the experimenter – is judged to be acceptable. However, there are also situations where the researcher wants to test which of two (or more) coindexation configurations is preferred. In this section, I consider methods that allow researchers to identify which is the "winning" antecedent for a particular anaphor in a particular syntactic or semantic configuration. Consider examples such as *John told Peter about the picture of him* or *Mary told Kate about the fountain near her* – who does the pronoun refer to? Crucially, one should not infer that a dispreferred interpretation is unacceptable/unavailable. It may be entirely acceptable but less preferred.

Prior work on coreference has used different means to convey the interpretations that participants have to choose between, including (i) asking participants to answer multiple-choice questions that present different

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

interpretation options, (ii) asking participants to choose between two pictures that depict two different antecedent choices, and (iii) asking participants to act out the meaning of the sentence using dolls (and thus indicating antecedent choice).

Let us first consider studies where participants answer multiple-choice questions, where the choices represent different interpretation options. One example comes from Kaiser and Runner's (2008) work on pronouns, reflexives, and emphatic forms in picture-NP constructions in German and Dutch, where participants read sentences like (15) (in Dutch) and answered a question about who is shown in the picture.

(15) a. Arne {vertelde/hoorde van} Hans over de foto van {hem/zichzelf/hemzelf}.
Arne {told/heard from} Hans about the picture of {pronoun/reflexive/emphatic pronoun}

The answer choices for (15) were (i) Arne (coded as subject), (ii) Hans (object), (iii) it could be either Arne or Hans, or (iv) someone else. Inclusions of the two final options allowed us to probe for level of ambiguity as well as potential configurations where sentence-external referents are strongly preferred. Our question wording allowed us to avoid metalinguistic notions like "refer to." Building on Kaiser et al. (2009), the verb type was manipulated (*tell/hear from*) as was the anaphoric form (*pronoun/reflexive*) in order to investigate how the distinction between sources (subject of *tell*, object of *hear from*) and perceivers (object of *tell*, subject of *hear from*) influences the interpretation of pronouns and reflexives.

In earlier work, Sturt (2003) also used multiple-choice questions but with only two answer choices, as illustrated in (15b). (These questions were part of a follow-up self-paced reading study.)

(15) b. Jonathan was pretty worried at the City Hospital. He remembered that the surgeon had pricked himself with a used syringe needle.
Who had been pricked with a used needle? Jonathan / the surgeon

Multiple-choice questions, but with more metalinguistic wording, were also used by Cunnings and Sturt (2014, 2018). They compared co-argument and non-co-argument configurations, as exemplified in (16a, b): In (16a), *him* is in a coargument configuration with *John*, but in the picture-NP structure in (16b), *him* is not a coargument of *John*. Cunnings and Sturt put a box around the anaphor in each sentence and asked people to "choose who they thought the boxed reflexive or pronoun most likely referred to" (Cunnings & Sturt 2014: 313). A multiple-choice format was used: in example (16), the choices were *Jonathan, the surgeon* or "either." (Cunnings and Sturt did not give their participants the option of selecting "someone else.")

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

(16) a. The surgeon remembered that Jonathan had noticed him near the back of the lunch queue.
b. The surgeon remembered about Jonathan's picture of him near the back of the lunch queue.

The "boxed pronoun" approach of Cunnings and Sturt (2014, 2018) is reminiscent of the earlier "circled pronoun" approach of Carden and Dieterich (1981): they investigated cataphora using sentences like (17a, b).

(17) a. The directors discussed the situation with Smith all afternoon. They finally decided that they would have to put him under the new Vice President, whether McIntosh liked it or not.
b. The directors discussed the situation with Smith all afternoon. They finally decided that he would have to report to the new Vice President, whether McIntosh liked it or not.

Here, an anaphoric candidate antecedent is available (*Smith*), but Carden and Dieterich note that real-world plausibility should bias *McIntosh*. One of the key research questions was how people interpret the object position pronoun *him* in (17a), given that a subject-position pronoun (in (17b), *he*) is more clearly judged to *not* be able to corefer with a subsequent name that it c-commands (Binding Principle C) – this question bears on different definitions of "command" (C-command vs. S-Command).

What is of interest to us, from a methodological perspective, is that Carden and Dieterich did not provide a list of answers to choose from, but instead instructed their participants to underline "the word to which the circled pronoun refers" (1981: 592). Thus, like Cunnings and Sturt (2014, 2018), they ask a metalinguistic question about reference, but did not provide people with a pre-existing set of answers to choose from. This can be advantageous, as it has been suggested that providing participants with specific lists of referents to choose from may artificially boost the salience of referents that would otherwise (normally) not be considered. (Underlining may have other drawbacks, e.g. unclear marks and manual data entry.)

Carden and Dieterich found that participants tend to interpret both subject and object pronouns as referring to *Smith* (anaphora) and not *McIntosh* (cataphora) – even though in other cataphoric conditions without any kind c-command participants were willing to interpret pronouns cataphorically – suggesting that the kind of structural superiority at play includes objects as well as subjects.

Another example of metalinguistic questions being used to probe referential interpretation comes from Patterson et al. (2014). In their study on native and non-native speakers' use of Binding Principle B, they used structures like (18) and asked participants to "read each sentence carefully and decide who the pronoun probably referred to" (2014: 4). Patterson et al. used the word "probably" to signal that another interpretation (a sentence-external referent) is also possible, but was not given as one of the multiple-choice options. After each sentence, participants saw the metalinguistic question “Who does [pronoun] refer to?” (see (18)) and had three options to choose from, as shown in (18).

(18) The boy remembered that Matthew had bought him a new computer game.
Who does “him” refer to?
The boy
Matthew
Either

So far, we have considered methodologies where the competing antecedent choices were presented linguistically/in writing. There are also studies where participants are asked to choose between two pictures depicting two different antecedent choices. One example is Sekerina et al.’s (2004) work on reflexives and short-distance pronouns (e.g. in locative constructions such as *behind herself/her*). This work relates to the broader question of whether pronouns and reflexives in locative prepositional phrases are in complementary distribution in terms of their antecedent choices. In one of their studies participants read preamble-question sequences like (19) and were presented with two pictures for each item: one with a sentence-internal referent for the anaphor (e.g. the box is behind the boy, and a man is standing nearby) and one with a sentence-external referent for the anaphor (e.g. the box is behind the man, and the boy is standing nearby).

(19) *Preamble:* In these pictures, you see a boy, a man, and a box. The boy has placed the box on the ground.
Question: Which picture shows that the boy has placed the box behind {himself/him}?
(a) the left picture
(b) the right picture
(c) both pictures

In another version, participants were presented with sentences like (20) and two images (one on the left and one on the right), and pressed a button to indicate their choice of the left or the right image (while their reaction times were recorded and eye movements tracked).

(20) Which picture shows that the boy has placed the box behind himself/him?

The results of both experiments reveal a non-complementarity in how pronouns and reflexives are interpreted: In their final responses, participants mostly choose sentence-internal referents for both reflexives and pronouns, but pronouns are more ambiguous than reflexives in also

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

allowing sentence-external referents in approx. 20 percent of the cases. This pattern is also reflected in eye movements. Reaction times were also longer for those pronouns trials where participants chose sentence-external referents. Sekerina et al. conclude that pronouns inside locative PPs are indeed ambiguous between a subject antecedent and a sentence-external antecedent.

In addition to methods where the antecedent possibilities that participants had to choose between were presented either linguistically or depicted visually, some researchers have opted for a combination of linguistic and visual presentation. An example of using both images and linguistic choices comes from Moulton et al. (2018). They investigated effects of focus on (potentially) cataphoric pronouns using sentences such as *His mother greeted Benny*, preceded by question contexts which either put the R-expression in focus (e.g. *His mother greeted which guy?*) or rendered it discourse-old and unfocused (e.g. "Who greeted him?") After hearing the question-answer sequence and seeing an image depicting four labeled characters (e.g. Benny, Benny's mother, Larry, Larry's mother), participants saw the question "The question you just heard was about Benny and WHO?" accompanied by images of Benny's mother and Larry's mother (labeled as such, so there was no memory burden). The task was to click with the mouse on the answer, which provides a measure of whether participants interpret the genitive *his/her* in the critical sentence as cataphoric (referring to *Benny*) or not. Thus, similar to Kaiser et al. (2009) and Kaiser and Runner (2008), this set-up allows the researchers to get a measure of participants' anaphor resolution without using metalinguistic terms such as "refer to."

It is also worth considering potential differences in short-term memory load induced by these methods. If the critical sentence is still visible when the question is presented, participants do not need to rely on a memory representation when answering the question. This is the case with Cunnings and Sturt (2014, Experiment 4) and Kaiser and Runner (2008), for example. If, however, the question is presented *after* the sentence has disappeared – as was the case in Sturt's (2003) follow-up study (as in (15b)), participants are answering based on their memory of the critical sentence. In this case, one may wonder if Binding-incompatible responses could be (partially) due to participants having to rely on a potentially "noisy" memory representation of the sentence.

In addition to these kinds of multiple-choice questions (where participants choose between written or visually presented options), some researchers have asked participants to act out the meaning of the critical sentence, thereby indicating how they interpret the anaphor. For example, Runner et al. (2003) researched the interpretation of pronouns and reflexives in possessed picture NPs (e.g. *Harry's picture of him/himself*). In possessed PNPs, according to standard Binding Theory, the reflexive must be, and the pronoun cannot be, coreferential with the possessor. To test participants'

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

interpretations of the anaphor inside the possessed PNP, Runner et al. used sequences like (21). Participants heard these while seated in front of a board with three male dolls seated in front of it (each doll was seated below three photographs, each showing one of the three dolls). The task was to act out the instruction provided in the second sentence. Participants' eye movements as well as their offline responses were recorded. Although harder to implement than some of the other methods, and not suitable for all linguistic configurations, the act-out task is very intuitive, has simple instructions (participants can simply be instructed to "do what they are told"), and does not involve any explicitly metalinguistic components.

(21) Look at Ken. Have Joe touch Harry's picture of himself.

Runner et al. found that contrary what traditional Binding Theory leads us to expect, (i) reflexives can be interpreted as coreferential with referents other than the possessor, and (ii) reflexives and pronouns are not in complementary distribution.

In sum, experimental work on coreference has also used a range of methods to probe choice of antecedent (Section 11.4) – differing in how the answer choices are presented and how metalinguistic the questions are – in addition to methods assessing acceptability of pre-specified coreference configurations (Section 11.2). The next section discusses the benefits of using both approaches in tandem.

11.4.1 Complementary Benefits of Acceptability Tasks and Antecedent-Choice Tasks

The methods described in the preceding section do not ask participants to assess the acceptability of a pre-specified coreference relation, but they can nevertheless provide useful complementary data for research that uses acceptability judgment tasks. This holds especially for non-co-argument contexts where reflexives and pronouns are not in fully complementary distribution, such as picture-NP contexts (e.g. *a picture of her/herself*, *a joke about him/himself*), locative structures (e.g. *near him/himself*) and other structures such as coordinations and comparatives (e.g. *linguists like her/herself*). Consider a situation where an anaphor can be coreferential with either of two possible antecedents, i.e. both coreference configurations are rated highly acceptable. Crucially, the antecedent-choice tasks described above could still reveal that one of the two candidate antecedents is preferred over the other, even though both are acceptable. Considering only the results of the acceptability judgment task would lead a researcher to overlook this difference.

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

Conversely, using only an antecedent-choice task could lead a researcher to observe that one of the candidate antecedents always wins over the other – which could lead a researcher to incorrectly conclude that the other candidate antecedent is unavailable and/or unacceptable. However, this conclusion may also be incorrect, as it could simply be that the other candidate antecedent is dispreferred (i.e. it "loses out" to the other competitor) even if it is fully acceptable.

In sum, the antecedent-choice tasks and acceptability rating tasks are best used in tandem and viewed as complementary approaches. Which is most suited depends on the hypothesis being tested, and often using both can yield valuable information.

11.5 Real-time Methods Used in Experiments on Anaphora

The offline methods we have discussed so far provide information about participants' final interpretation or assessment of a coreference relation. They do not provide direct information about the time-course of people's decision-making, about the load that the response placed on the processing system, or possible transient processing effects that occurred before people reached their final decision. This kind of information can be obtained by means of online methods, including (i) eye-tracking during reading, (ii) visual-world eye-tracking, and (iii) self-paced reading. Neurolinguistic methods such as ERP have also been used (see e.g. Harris et al. (2000), Xiang et al. (2009) on neurolinguistic investigations of coreference). These online methods allow researchers to obtain information about processing that occurs before a comprehender reaches their "final decision" about what an anaphor refers to (see Chapters 22–24 for further discussion).

To see how the three most widely used methods can provide information about real-time processing, I review an example of each. In self-paced reading, participants read sentences word-by-word (or chunk-by-chunk) and the reading time for each word is measured. In the "moving window" variant of self-paced reading, preceding and upcoming words are masked. Using this method, Badecker and Straub (2002) investigated sentences like (22) to see whether, when participants encounter the reflexive *himself* or the pronoun *him*, they only consider the structurally licensed antecedent (the local subject *Bill* in the case of *himself*, and the non-local subject *John* in the case of *him*) or whether people's reading times are sensitive to the presence of a gender-matching but structurally inaccessible candidate referent. If Binding constraints "kick in" right away, the presence/absence of a gender-matching, structurally inaccessible referent should have no effect:

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

(22) a. {Jane/John} thought that Bill owed himself another opportunity to solve the problem.
b. John thought that {Bill/Beth} owed him another opportunity to solve the problem.

With reflexives (22a), the presence of a gender-matching matrix subject should not, according to Binding Theory, influence the processing of the reflexive because the matrix subject is structurally inaccessible. Similarly, with pronouns (22b), the presence of a gender-matching embedded subject should not influence processing of the pronoun, because the local subject is not, according to Binding Theory, a potential antecedent for the pronoun.

However, Badecker and Straub (2002) found that both pronouns and reflexives in sentences with two gender-matching "candidate antecedents" were read slower than sentences with no gender-matching inaccessible antecedent – suggesting competition from the supposedly inaccessible candidate. Badecker and Straub conclude that during real-time processing, comprehenders initially consider all sufficiently salient referents, regardless of whether they meet the structural requirements for coreference. These slowdowns are an example of online methods being able to pick up on transient processes that can inform theories of reference resolution. However, there is an ongoing debate about who exactly gets to compete in the processing of anaphoric dependencies.

Eye-tracking during reading offers a more fine-grained way of measuring reading time, because in this paradigm, readers can freely move through the text, including back-tracking and rereading. Thus, instead of a single reaction time per word, as is the case with self-paced reading, self-paced reading allows researchers to obtain multiple measures, such as (i) first-fixation duration (the duration of the first fixation on a particular region), (ii) first-pass reading time (the sum of all the fixation durations in a region from when a reader first enters it to when they first leave the region), and (iii) second-pass reading time (the sum of all fixation durations on a region after that region has already been exited once). Sturt (2003) used eye-tracking during reading to investigate the same question as Badecker and Straub (2002), using sentences involving stereotypically gendered nouns (e.g. *surgeon*), as shown in (23):

(23) Jonathan/Jennifer was pretty worried at the City Hospital. He/She remembered that the surgeon had pricked himself/herself with a used syringe needle. There should be an investigation soon.

Sturt found no effects of the structurally inaccessible character on first-fixation times, but he did find effects of the inaccessible character on second-pass reading times. This led him to conclude that although the structurally inaccessible referent has an effect during the later stages of processing, Binding Theory constraints nevertheless kick in very early on,

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

contrary to the conclusions of Badecker and Straub. The availability of both early and late reading time measures means that eye-tracking during reading can provide a more detailed look at online processing than self-paced reading. Relatedly, work on the processing of cataphora (e.g. Drummer & Felser (2018) with eye-tracking during reading, Kazanina et al. (2007) with self-paced reading) has led to divergent results, which Drummer and Felser (2018) suggest may be due to the less fine-grained timing information available from self-paced reading.

In addition to methods that measure *processing time* (e.g. self-paced reading and eye-tracking during reading), other online methods such as visual-world eye-tracking offer a means of probing *antecedent activation* over time. For example, in Runner et al.'s (2003) study, described above, participants heard sentences like "Pick up Joe. Look at Ken. Have Joe touch Harry's picture of himself/him." Runner et al. tested picture-NPs with possessors – a structure where, according to Binding Theory, reflexives should always be interpreted as coreferential with the possessor and pronouns should never receive this interpretation. Participants were seated at a board with three male dolls (Ken, Joe, Harry) sitting in front of it, and each doll was seated below three photographs (one of each of the dolls). The participants acted out the instruction, thereby indicating their offline response, while their eye movements to the different dolls and pictures were recorded, thereby providing a measure of what participants are considering as potential antecedents (and not a measure of processing slowdowns or processing load).

Crucially, the eye movements in the reflexive conditions show no evidence that people are looking at the possessor doll (e.g. Harry) earlier than the subject doll (e.g. Joe) – in other words, Runner et al. find no evidence of looks to "structurally accessible" (Binding-Theory-compatible) referents preceding looks to "structurally inaccessible" (Binding-Theory-incompatible) referents. The information about what potential antecedents participants consider at different points in time is crucial for Runner et al. to argue against a view of Binding Theory as an "initial filter" that constrains the earliest moments of processing (e.g. Nicol & Swinney 1989). Offline data about the proportion of antecedent choices could not be used to shed light on claims regarding the timing of when Binding principles kick in. That being said, it is important in this paradigm to collect *both* offline and online data. Because Runner et al. also had data about offline choices, they were able to conduct a follow-up analysis looking at eye movements in only those trials where participants chose the possessor as the antecedent of the reflexive (i.e. made the doll touch the picture of Harry) – and they find that even on these trials, where participants' offline responses are in line with standard Binding Theory, there is no sign of people looking at the Binding-Theory-compatible possessor earlier than the non-Binding-Theory-compatible subject referent. In this way, online and offline data can be used to complement each other.

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

The modality of stimulus presentation is different in visual-world eye-tracking and eye-tracking during reading: auditory vs. written. Thus, visual-world eye-tracking can be used to investigate anaphora processing in preliterate children, and can also be used to investigate prosodic effects. (Conversely, when recording stimuli for visual-world experiments, it is important to control the prosodic properties of the stimuli to avoid inadvertent confounds such as certain pitch accent patterns present in some conditions but not in others.)

It is worth noting that the three online methods discussed here typically do not require a metalinguistic task – i.e. participants are typically not asked to judge whether a sentence is acceptable or whether a certain word can refer to the same entity as another word – in contrast to many of the acceptability methods described earlier in this chapter.

11.6 Empirical and Theoretical Results

A full discussion of the empirical and theoretical results of experimental work on coreference is beyond the scope of this chapter. However, in this section I will briefly review some of the key contributions and ongoing debates that have come out of experimental work on coreference.

As has already become clear over the course of this chapter, experimental approaches using offline methods have contributed significantly especially in areas where intuitions tend to be murky. We have already mentioned experimental work using acceptability-judgment tasks and related methods that investigated structures such as picture-NPs (e.g. Keller & Asudeh 2001; Kaiser & Runner 2009), locative PPs (e.g. Sekerina et al. 2004), anaphora inside coordinations (Foraker 2003), as well as cataphoric configurations (e.g. Moulton et al. 2018; Temme & Verhoeven 2017). The results of these studies help to clarify the empirical landscape, thereby strengthening syntactic theorizing and our understanding of the relation between syntactic, semantic, and pragmatic information.

In addition to helping to clarify theoretically significant judgments, experimental work on coreference contributes to our understanding of the memory representations and processes involved in the interpretation of dependencies. Traditionally, the question of whether comprehenders only consider structurally licensed, Binding-Theory-compatible antecedents during online processing of anaphora, or whether feature-matching but structurally inaccessible referents are also temporarily activated as possible antecedents emerged as one of the key questions in the late 1980s. Independently of how exactly the principles of Binding Theory are formulated in different syntactic frameworks, a key question is whether or not these principles fully constrain language processing from the earliest moments onwards.

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

Before getting into the processing details of this question, it is worth emphasizing that one first needs to reliably establish what are the structural constraints that guide comprehenders' ultimate choices about anaphoric coreference. In order to see when and to what extent real-time processing is guided by structural considerations, we need to know what the relevant structural factors are. This is relatively well researched for languages like English: Linguists' intuitions about core Binding Theory configurations have been (largely) confirmed by experimental work such as Gordon and Hendrick (1997). There also exists some foundational experimental work in areas where the complementarity of pronouns and reflexives is less clear, such as picture-NPs (*picture of her/herself*) and locative PPS (*behind her/herself*). However, if one wants to investigate the time course of Binding constraints in a less well-researched language, then offline, acceptability-based investigations would presumably be conducted beforehand or simultaneously with online processing studies. Even if one's ultimate aim is an investigation of real-time anaphoric processing, the offline methodologies described in the earlier parts of this chapter can be used to provide the backdrop that is needed for interpreting the online data.

The online status of Binding Theory has far-reaching consequences for our understanding of the mechanisms involved in anaphor resolution and the construction of linguistic dependencies more generally. If we regard anaphor resolution as a memory search problem, the question becomes: When faced with an anaphoric element, comprehenders have to search through memory to find its antecedent. Is this memory search/retrieval process subject to interference from structurally inaccessible referents that match the phi-features of the anaphoric expression? Or is it purely structurally guided and insensitive to non-structural cues such as gender, person, and number? (See e.g. Dillon (2014) for a recent overview.) Furthermore, given that language involves many other kinds of dependencies as well (e.g. subject–verb agreement, licensing of negative polarity items, etc.), one would like to know whether the retrieval process involved with anaphora is similar to the retrieval processes involved in these other kinds of dependencies. To be able to answer these questions, we need to be able to tap into the incremental, moment-by-moment processes that comprehenders engage in after encountering an anaphoric expression.

Early work by Nicol and Swinney (1989) argued in favor of an initial-filter approach, according to which Binding Theory constrains processing immediately and successfully. According to this kind of view, reflexive dependencies are not subject to interference from structurally inaccessible elements. However, Badecker and Straub (2002) – discussed above – found evidence that the processing of both pronouns and reflexives involved competition between structurally accessible and inaccessible referents. But this finding has not been consistently replicated for object-

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

position reflexives: subsequent findings are mixed (e.g. Sturt 2003; Cunnings & Felser 2013; Dillon et al. 2013; Cunnings & Sturt 2014, 2018; Patil et al., 2016). Looking at non-object-position reflexives, Runner et al.'s (2003) and Kaiser et al.'s (2009) work on possessed picture-NPs found no evidence for early effects of Binding Theory.

Further complicating the picture,[10] in recent work, Parker et al. (2015) tested null subject anaphora (in configurations where they are often analyzed as PRO by syntacticians) and found that while overt reflexive pronouns seem impervious to interference effects, the resolution of null PRO subjects is susceptible to interference (thereby resembling other kinds of dependency formation processes such as subject–verb agreement and the licensing of negative polarity items; see e.g. Pearlmutter et al. 1999; Vasishth et al. 2008; Xiang et al. 2009; Wagers et al. 2009). This finding leads Parker et al. to conclude that different kinds of anaphora may differ in their susceptibility to interference (see also Parker & Phillips 2017). This is an active area for future work.

11.7 Open Questions and Future Directions

Recent years have seen an increase in experimental work related to anaphor resolution that had generated fruitful empirical and theoretical insights. Nevertheless, many questions are still open and many phenomena and languages remain under-investigated.

In terms of cross-linguistic research, prior experimental work on anaphora in languages including Mandarin, Korean, German, and Dutch (e.g. Kaiser & Runner 2008; Schumacher et al. 2011; Dillon et al. 2014; Han et al. 2015; Dillon et al. 2016; He & Kaiser 2016) has already significantly broadened our understanding of how anaphora resolutions work in morphologically diverse systems. However, further broadening the empirical domain of investigation beyond commonly researched (and often typologically similar) languages is important, if our aim is to gain an understanding of human language (see also Chapter 7). It's also worth noting that phenomena that may at first appear to be limited in scope may be present in other languages as well, once we know how to look for them (see e.g. Sloggett & Dillon's (2018) finding that English long-distance reflexives exhibit person blocking effects similar to Mandarin). Thus, increased research on less-researched languages also has the potential to improve our understanding of more commonly researched languages. In general, working on under-researched languages is a domain where one can clearly see the utility of simple, offline methods that can be utilized in settings outside of the typical university laboratories but that nevertheless allow us to get a sense of the amount of variability vs. stability in the data.

[10] See also Nicenboim et al. (2018) on the small magnitude of interference effects the domain of agreement, which hints at the possibility that lack of power may be preventing the detecting of interference effects.

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

In terms of linguistic phenomena, there exist various phenomena having to do with reference that are known in theoretical linguistics but have received little or no attention in experimental work. Consider, for example, the pronouns in examples (24a–c) below (from Büring 2011). These are all pronouns that *cannot* be interpreted as "pointing to" a specific previously mentioned (or upcoming) entity. Thus, they differ from the majority of the cases that are typically investigated in psycholinguistic work and raise intriguing questions for current psycholinguistic models of anaphora.

(24) a. This year the president is a Republican, but one fine day, he will be a member of the Green Party.
b. Mary, who deposited her paycheck at the ATM, was smarter than any woman who kept it in her purse.
c. Every farmer who owned a donkey had Lucy vaccinate it.

In (24a) the pronoun *he* does *not* refer to the actual referent of the noun *the president* in the preceding clause. In other words, (24a) does not mean that the person who is the current president will one day become a member of the Green Party. Instead, *he* denotes a function that picks out the individual who is the president at the relevant time in the relevant world (see e.g. Büring (2011) for an overview). This means that the pronoun in (24a) cannot be interpreted as being straightforwardly coreferential with what might at first blush appear to be its antecedent, i.e. the referent of *the president* in the preceding clause. These kinds of pronouns are called *pronouns of laziness* (Geach 1962). A related challenge is posed by examples like (24b). Here, the pronoun *it* does not refer to Mary's paycheck. Instead it can be thought of as meaning '*her* paycheck' where *her* is a bound pronoun bound by *any woman*. Thus, similar to (24a), the relation between the pronoun *it* and its antecedent is more complex than with regular referential pronouns. Pronouns like *it* in (24b) are called *paycheck pronouns* (Karttunen 1969). A third class are *donkey pronouns or E-type pronouns* (Evans 1977), illustrated in (24c). In (24c), *it* does not point to a specific donkey – instead, the sentence means something like 'Every farmer who owned a donkey had Lucy vaccinate the donkey owned by *him*'. So the pronoun *it* acts like a definite description containing a pronoun (*him*) that is bound by *every farmer*.

All of these examples are grammatical and comprehensible, but they involve pronouns that are not simply pointers to an already mentioned or upcoming referent. Thus, they differ from sentences like *Lisa was tired. Joan helped her* or *Joan helped herself*, where we can think of the anaphoric expression as pointing to a specific antecedent (*Lisa* or *Joan*). There exists extensive theoretical literature one each of these three pronoun types, but their consequences for psycholinguistic models of reference remain largely underexplored (but see e.g. Grosz et al. (2015), Kush & Eik (2019) for experimental work on donkey anaphora). Other reference-related topics, such as reciprocal pronouns (*each other*, *one another*) and locative constructions (e.g.

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

behind her/herself, also known as “snake sentences”) are known to exhibit some unusual binding properties in English as well as considerable cross-linguistic variation, but have received relatively little attention in the experimental literature.

Broadly speaking, using experimental methods to investigate reference-related phenomena that have not previously been approached from a psycholinguistic perspective can have at least two advantages (depending on the topic): first, in some cases the judgments can be murky, so experimental methods would allow researchers to obtain additional data and to assess the extent of inter- and intraspeaker variance. From this point of view, experimental syntax can help clarify the empirical bases of formal linguistic theories on anaphora.

Second, many of these phenomena can potentially help contribute to central debates in psycholinguistics, such as the questions regarding the kinds of retrieval mechanisms that are involved in interpreting anaphoric dependencies.

In conclusion, although the experimental investigation of coreference – in particular the use of acceptability judgments without recourse to subscripts – requires one to tread carefully when implementing one’s experiment, the resulting insights can contribute to work in both theoretical and experimental syntax.

References

Bach, E. & Partee, B. H. (1980). Anaphora and semantic structure. In J. Kreiman & A. E. Ojeda, eds., *Papers from the Parasession on Pronouns and Anaphora*. Chicago: Chicago Linguistic Society, pp. 1–28. [Reprinted in Partee, B. H. 2004. *Compositionality in Formal Semantics: Selected Papers by Barbara H. Partee*. Oxford: Blackwell, 122–152.]

Badecker, W. & Straub, K. (2002). The processing role of structural constraints on interpretation of pronouns and anaphors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(4), 748–769.

Bader, M. & Häussler, J. (2010). Toward a model of grammaticality judgments. *Journal of Linguistics*, 46, 273–330.

Bard, E. G., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, 72, 32–68.

Bornkessel-Schlesewsky, I. & Schlesewsky, M. (2007). The wolf in sheep’s clothing: against a new judgment-driven imperialism. *Theoretical Linguistics*, 33, 319–333.

Büring, D. (2011). Pronouns. In K. von Heusinger, C. Maienborn, & P. Portner, eds., *Semantics: An International Handbook of Natural Language Meaning* (Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science 33/2). Berlin: De Gruyter Mouton, pp. 971–996.

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

Carden, G. & Dieterich, T. (1981). Introspection, observation, and experiment: An example where experiments pay off. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, 1980*. Chicago: University of Chicago Press, pp. 583–597.

Carminati, M.-N. (2005). Processing reflexes of the Feature Hierarchy (Person > Number > Gender) and implications for linguistic theory. *Lingua*, 115, 259–285.

Carminati, M.-N., Frazier, L., & Rayner, K. (2002). Bound variables and c-command. *Journal of Semantics*, 19, 1–34.

Chien, Y.-C. & Wexler, K. (1990). Children's knowledge of locality conditions in binding as evidence for the modularity of syntax and pragmatics. *Language Acquisition*, 1, 225–295.

Cowart, W. (1997). *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. Thousand Oaks, CA: Sage.

Culbertson, J. & Gross, S. (2009). Are linguists better subjects? *British Journal for the Philosophy of Science*, 60, 721–736.

Cunnings, I. & Felser, C. (2013). The role of working memory in the processing of reflexives. *Language and Cognitive Processes*, 28, 188–219.

Cunnings, I., Patterson, C., & Felser C. (2015) Structural constraints on pronoun binding and coreference: Evidence from eye movements during reading. *Frontiers in Psychology*, 6, 840.

Cunnings, I. & Sturt, P. (2014). Coargumenthood and the processing of reflexives. *Journal of Memory and Language* 75, 117–139.

Cunnings, I. & Sturt, P. (2018). Coargumenthood and the processing of pronouns. *Language, Cognition and Neuroscience*, 33(10), 1235–1251.

Dąbrowska, E. (2010). Naïve v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review*, 27, 1–23.

Dillon, B. (2014). Syntactic memory in the comprehension of reflexive dependencies: An overview. *Language and Linguistics Compass*, 8(5), 171–187.

Dillon, B., Chow W.-Y., Wagers, M., Guo, T., Liu, F., & Phillips, C. (2014). The structure-sensitivity of memory access: Evidence from Mandarin Chinese. *Frontiers in Psychology*, 5, 1025.

Dillon, B., Chow, W.-Y., & Xiang, M. (2016). The relationship between anaphor features and antecedent retrieval: Comparing Mandarin *ziji* and *ta-ziji*. *Frontiers in Psychology*, 6, 1966.

Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, 69, 85–103.

Drummer, J.-D. & Felser, C. (2018). Cataphoric pronoun resolution in native and non-native sentence comprehension. *Journal of Memory and Language*, 101, 97–113.

Evans, G. (1977). Pronouns, quantifiers, and relative clauses. *The Canadian Journal of Philosophy*, 7(3), 467–536.

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

Fabrigar, L. R. & Paik, J.-E. S. (2007). Thurstone scales. In N. Salkind, ed., *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage, pp. 1003–1005.

Featherston, S. (2002). Coreferential objects in German: Experimental evidence on reflexivity. *Linguistische Berichte*, 192, 457–484.

Featherston, S. (2008). Thermometer judgments as linguistic evidence. In C. M. Riehl & A. Rothe, eds., *Was ist linguistische Evidenz?* Aachen: Shaker, pp. 69–90.

Foraker, S. (2003). The processing of logophoric reflexives shows discourse and locality constraints. Paper presented at the 38th Annual Meeting of the Chicago Linguistic Society.

Fraundorf, S. H., Benjamin, A. S., & Watson, D. G. (2013). What happened (and what didn't): Discourse constraints on encoding of plausible alternatives. *Journal of Memory and Language*, 69, 196–227.

Frazier, L. & Clifton, C. (2000). On bound variable interpretations: The LF-Only Hypothesis. *Journal of Psycholinguistic Research*, 29, 125–139.

Fukuda, S., Goodall, G., Michel, D., & Beecher, H. (2012). Is Magnitude Estimation worth the trouble? In J. Choi, E. A. Hogue, J. Punske, D. Tat, J. Schertz, & A. Trueman, eds., *Proceedings of the 29th West Coast Conference on Formal Linguistics*. Somerville, MA: Cascadilla Proceedings Project, pp. 328–336.

Fukumura K., Hyönä, J., & Scholfield, M. (2013). Gender affects semantic competition: The effect of gender in a non-gender-marking language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(4), pp. 1012–1021.

Garnham, A. (2001). *Mental Models and the Interpretation of Anaphora*. Hove: Psychology Press.

Geach, P. (1962). *Reference and Generality*. Ithaca, NY: Cornell University Press.

Gerken, L.-A. & Bever, T. (1986). Linguistic intuitions are the result of interactions between perceptual processes and linguistic universals. *Cognitive Science*, 10, 457–476.

Gibson, E. & Fedorenko, E. (2013). The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, 28(1–2), 88–124.

Gordon, P. C. & Hendrick, R. (1997). Intuitive knowledge of linguistic co-reference. *Cognition*, 62, 325–370.

Grosz, P. G., Patel-Grosz, P., Fedorenko, E., & Gibson, E. (2015). Constraints on donkey pronouns. *Journal of Semantics*, 32(4), 619–648.

Han, C.-h., Storoshenko, D., Leung, B., & Kim, K. (2015). The time course of long distance anaphor processing in Korean. *Korean Linguistics*, 17 (1), 1–32.

Harris, T., Wexler, K., & Holcomb, P. (2000). An ERP investigation of binding and coreference. *Brain and Language*, 75, 313–346.

Häussler, J. & Juzek, T. S. (2017). Hot topics surrounding acceptability judgement tasks. In S. Featherston, R. Hörnig, R. Steinberg, B. Umbreit,

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

& J. Wallis, eds., *Proceedings of Linguistic Evidence 2016: Empirical, Theoretical, and Computational Perspectives*. University of Tübingen. http://dx.doi.org/10.15496/publikation-19039

He, X. & Kaiser, E. (2016). Processing the Chinese reflexive “ziji”: Effects of featural constraints on anaphor resolution. *Frontiers in Psychology*, 7, 284.

Huang Y. (2000). *Anaphora: A Cross-Linguistic Approach*. Oxford: Oxford University Press.

Ionin, T. & Zyzik, E. (2014). Judgment and interpretation tasks in second language research. Review article for *Annual Review of Applied Linguistics*. *Annual Review of Applied Linguistics*, 34, 1–28.

Kaiser, E. (2015). Perspective-shifting and free indirect discourse: Experimental investigations. In S. D’Antonio, M. Moroney, & C. R. Little, eds, *Proceedings of Semantics and Linguistic Theory 25 (SALT 25)*, pp. 346–372.

Kaiser, E., Nichols, J., & Wang, C. (2018). Experimenting with imposters: What modulates choice of person agreement in pronouns? *Proceedings of Sinn und Bedeutung*, 22(1), 505–521.

Kaiser, E. & Runner, J. T. (2008). Intensifiers in German and Dutch Anaphor Resolution. In N. Abner & J. Bishop, eds., *Proceedings of the 27th West Coast Conference on Formal Linguistics*. Somerville, MA: Cascadilla Proceedings Project, pp. 265–273.

Kaiser, E., Runner, J., Sussman, R., & Tanenhaus, M. (2009). Structural and semantic constraints on the resolution of pronouns and reflexives. *Cognition*, 112, 55–80.

Karttunen, L. (1969). Pronouns and variables. In R. I. Binnick, A. Davidson, G. M. Green, & J. L. Morgan, eds., *Proceedings of the Fifth Regional Meeting of the Chicago Linguistic Society*. Chicago: Chicago Linguistic Society, pp. 108–116.

Kazanina, N., Lau, E. F., Lieberman, M., Yoshida, M., & Philips, C. (2007). The effect of syntactic constraints on the processing of backwards anaphora. *Journal of Memory and Language*, 56, 384–409.

Keller, F. & Asudeh, A. (2001). Constraints on linguistic coreference: Structural vs. pragmatic factors. In J. Moore & K. Stenning (eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum, pp. 483–488.

Keller, F. (2000). Gradience in grammar: Experimental and computational aspects of degrees of grammaticality. Doctoral dissertation, University of Edinburgh.

Kreiner, H., Sturt, P., & Garrod, S. (2008). Processing definitional and stereotypical gender in reference resolution: Evidence from eye-movements. *Journal of Memory and Language*, 58(2), 239–261.

Kush, D. & Eik, R. (2019). Antecedent accessibility and exceptional covariation: Evidence from Norwegian Donkey Pronouns. *Glossa: A Journal of General Linguistics*, 4(1), 96. DOI:10.5334/gjgl.930

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

Kush, D., Lidz, J., & Phillips, C. (2015). Relation-sensitive retrieval: Evidence from bound variable pronouns. *Journal of Memory and Language*, 82, 18–40.

Labov, W. (1972). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania.

Lago, S., Stutter Garcia, A., & Felser, C. (2018). The role of native and non-native grammars in the comprehension of possessive pronouns. *Second Language Research*, 35(3), 319–349.

Langsford, S. et al. (2018). Quantifying sentence acceptability measures: Reliability, bias, and variability. *Glossa: A Journal of General Linguistics*, 3(1), 37. DOI:10.5334/gjgl.396

Maia, J. & Morris, R. (2019). The semantics–pragmatics of typographic emphasis in discourse. Poster presented at the 32nd Annual CUNY Conference on Human Sentence Processing.

Montrul, S., Dias, R., & Santos, H. (2011). Clitics and object expression in the L3 acquisition of Brazilian Portuguese: Structural similarity matters for transfer. *Second Language Research*, 27, 21–58.

Moulton, K., Chan, Q., Cheng, T., Han, C.-h., Kim, K., & Nickel-Thompson, S. (2018). Focus on cataphora: Experiments in context. *Linguistic Inquiry* 49 (1) 151–168.

Myers, J. (2009). Syntactic judgment experiments. *Language and Linguistics Compass*, 3, 406–423.

Nicenboim, B., Vasishth, S., Engelmann, F., & Suckow, K. (2018). Exploratory and confirmatory analyses in sentence processing: A case study of number interference in German. *Cognitive Science*, 42, 4, 1075–1100.

Nicol, J. & Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *Journal of Psycholinguistic Research*, 18, 5–20.

Parker, D. & Phillips, C. (2017). Reflexive attraction in comprehension is selective. *Journal of Memory and Language*, 94, 272–290.

Parker, D., Lago, S., & Phillips, C. (2015). Interference in the processing of adjunct control. *Frontiers in Psychology*, 6, 1–13.

Patil, U., Vasishth, S., & Lewis, R. L. (2016). Retrieval interference in syntactic processing: The case of reflexive binding in English. *Frontiers in Psychology*, 7, 1–18.

Patterson C., Trompelt, H., & Felser, C. (2014). The online application of binding condition B in native and non-native pronoun resolution. *Frontiers in Psychology*, 5, 147.

Pearlmutter, N., Garnsey, S., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and Language*, 41, 427–456.

Phillips, C. & Lasnik, H. (2003). Linguistics and empirical evidence: Reply to Edelman and Christiansen. *Trends in Cognitive Sciences*, 7, 61–62.

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

Pollard, C. & Sag, I. (1992). The processing of logophoric reflexives shows discourse and locality constraints. *Linguistic Inquiry*, 23(2), 261–303.

Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning*, 63, 595–626.

Reinhart, T. (1982). Pragmatics and linguistics: An analysis of sentence topics. *Philosophica*, 27, 53–94.

Reinhart, T. (1983a). *Anaphora and Semantic Interpretation*. London: Croom Helm.

Reinhart, T. (1983b). Coreference and bound anaphora: A restatement of the anaphora question. *Linguistics and Philosophy*, 6, 47–88.

Reinhart, T. (2002). The theta system: An overview. *Theoretical Linguistics*, 28, 229–290.

Rigalleau, F., Caplan, D., & Baudiffier, V. (2004). New arguments in favour of an automatic gender pronominal process. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 57A(5), 893–933.

Roberts, J., Laughlin, J., & Wedel, D. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement*, 59(2), 211–233.

Runner, J. T., Sussman, R. S., & Tanenhaus, M. K. (2003). Assignment of reference to reflexives and pronouns in picture noun phrases: Evidence from eye movements. *Cognition*, 89, B1–B13.

Schumacher, P. B., Bisang, W., & Sun, L. (2011). Perspective in the processing of the Chinese reflexive *ziji*: ERP evidence. *Lecture Notes in Computer Science*, 7009, 119–131.

Schütze, C. (1996). *The Empirical Base of Linguistics*. Chicago: University of Chicago Press.

Schütze, C. & Sprouse, J. (2013). Judgment data. In R. J. Podesva & D. Sharma, eds., *Research Methods in Linguistics*. Cambridge: Cambridge University Press, pp. 27–50.

Sekerina, I., Stromswold, K., & Hestvik, A. (2004). How adults and children process referentially ambiguous pronouns. *Journal of Child Language*, 31, 123–152.

Sloggett, S. & Dillon, B. (2018). Person blocking in reflexive processing: When “I” matter more than “them.” Talk given at CUNY 2018, UC Davis.

Sprouse, J. (2007). Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics*, 1, 123–134.

Sprouse, J. & Almeida, D. (2017). Design sensitivity and statistical power in acceptability judgment experiments. *Glossa: A Journal of General Linguistics*, 2(1), 14.1–32. DOI:10.5334/gjgl.236

Sprouse, J., Schütze, C. T., & Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from *Linguistic Inquiry* 2001–2010. *Lingua*, 134, 219–248.

Stevens, S. (1975). *Psychophysics: Introduction to Its Perceptual, Neural, and Social Prospects*. New York: John Wiley.

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, 48, 542–562.

Temme, A. & Verhoeven, E. (2017). Backward binding as a psych effect: A binding illusion? *Zeitschrift für Sprachwissenschaft*, 36(2), 279–308.

Thurstone, L. (1927). A law of comparative judgment. *Psychological Review*, 34(4), 273.

Vasishth, S., Brüssow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, 32(4), 685–712.

Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, 61(2), 206–237.

Weskott, T. & Fanselow, G. (2011). On the informativity of different measures of linguistic acceptability. *Language*, 87(2), 249–273.

Xiang, M., Dillon, B., & Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain and Language*, 108, 1, 40–55.

https://doi.org/10.1017/9781108569620.012 Published online by Cambridge University Press