Antecedent frequency effects during the processing of pronouns

Roger P.G. van Gompel a,*, Asifa Majid b

a Department of Psychology, University of Dundee, Dundee DD1 4HN, Scotland, UK
b Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Received 13 June 2003; accepted 10 July 2003

Abstract

An eye-movement reading experiment investigated whether the ease with which pronouns are processed is affected by the lexical frequency of their antecedent. Reading times following pronouns with infrequent antecedents were faster than following pronouns with frequent antecedents. We argue that this is consistent with a saliency account, according to which infrequent antecedents are more salient than frequent antecedents. The results are not predicted by accounts which claim that readers access all or part of the lexical properties of the antecedent during the processing of pronouns.

© 2004 Elsevier B.V. All rights reserved.

Keywords: Pronouns; Anaphors; Antecedent reaccess; Lexical frequency; Saliency; Focus; Lemma; Lexeme; Eye movements

1. Introduction

Experimental studies investigating pronoun resolution have shown that there are many linguistic factors that affect the ease with which pronouns are processed. Most studies have focussed on how high-level linguistic factors such as discourse information, sentence focus, and semantics affect pronoun resolution (see Garnham, 2001, for an overview). In contrast, very few studies have investigated whether and how lower-level factors related to the lexical properties of the antecedent influence pronominal processing. In this article, we will report an eye-movement experiment that investigated the effect on pronominal processing of one such low-level factor, namely the lexical frequency of a pronoun’s

* Corresponding author. Tel.: +44-1382-348117; fax: +44-1382-229993.
E-mail address: r.p.g.vangompel@dundee.ac.uk (R.P.G. van Gompel).
antecedent. The question that this experiment addressed is: Does the ease with which a pronoun (e.g. *his* in [1]) is processed depend on whether the lexical frequency of the antecedent is high (*criminal*) or low (*arsonist*)?

1. The constables suspected the criminal/arsonist of the break-in. They searched in his bag for evidence.

There are three different ways lexical frequency of the antecedent could affect pronoun resolution. When readers process a pronoun, they need to reaccess at least some information about the antecedent in order to establish a coreference link with it. One possible hypothesis is that reaccessing antecedents is very similar to lexical access for words that are not anaphoric; when readers process a pronoun, they retrieve the same lexical information about the preceding antecedent as they retrieved when they first encountered the antecedent noun. We will term this hypothesis the *full reaccess hypothesis*.

Indeed, a number of studies suggest that, at least in some respects, reaccessing antecedents is very similar to lexical access. Shillcock (1982) showed that after hearing a pronoun, lexical decision times to words that were semantically related to the antecedent were facilitated relative to words that were unrelated. This effect is similar to the semantic priming effect that is observed in lexical processing (e.g. Meyer & Schvaneveldt, 1971). Furthermore, there is considerable evidence that the processing of pronouns is sensitive to the syntactic gender of their antecedent in Romance languages (e.g. Cacciari, Carreiras, & Barbolini Cionini, 1997; Garnham, Oakhill, Ehrlich, & Carreiras, 1995), suggesting that readers reaccess the gender of a pronoun’s antecedent as they do during lexical access.

Because numerous studies have found robust frequency effects during lexical access (e.g. Forster & Chambers, 1973; Rayner & Duffy, 1986; Schilling, Rayner, & Chumbley, 1998; Whaley, 1978), the full reaccess hypothesis predicts that antecedent frequency effects should be observed at pronouns. When reading a pronoun, infrequent antecedents should be harder to reaccess than frequent antecedents, resulting in longer reading times for the pronoun.

A second account, endorsed by Simner and Smyth (1999), does not accept that reaccessing antecedents is the same as lexical access. In line with models of language production (e.g. Kempen & Huijbers, 1983; Levelt, 1989), they make a distinction between a word’s lemma and its lexeme. Semantic information and syntactic features (e.g. syntactic gender) are thought to belong to a word’s lemma, whereas word frequency is considered to be part of a word’s lexeme (Harley & MacAndrew, 2001; Jescheniak & Levelt, 1994; but cf. Dell, 1990). Simner and Smyth (1999) argued that the processor reaccesses lemma information when it encounters an anaphor, but not lexeme information. We will therefore call this the *lemma reaccess account* (see Jescheniak, Schriefers, & Hantsch, 2001; Meyer & Boeck, 1999, for similar arguments about pronoun production, but cf. Schmitt, Meyer, & Levelt, 1999). Because pronouns reaccess an antecedent’s lemma, the lemma reaccess account is consistent with the earlier-mentioned findings that pronouns reactivate their antecedent’s semantic information and syntactic gender. However, with respect to antecedent frequency effects, pronouns with infrequent antecedents should be no harder or easier to read than pronouns with frequent antecedents because word frequency is part of an antecedent’s lexeme.
Finally, Pynte and Colonna (2000) have suggested a third way in which antecedent frequency could affect processing of pronouns. According to the saliency account, lexical frequency affects the saliency of a word such that infrequent words are more salient than frequent words. And so, infrequent antecedents should be easier to access than frequent antecedents. Indeed, in the memory literature, it has been argued that infrequent words attract more attention than frequent words, and therefore are encoded into memory more strongly. Evidence for this comes from studies which show that recognition of previously presented infrequent words is better than recognition of frequent words (e.g. Glanzer & Adams, 1985, 1990; Malmberg & Nelson, 2003). Furthermore, there is evidence that, at least under some circumstances, linguistic material that requires more processing effort is better encoded into memory (e.g. McDaniel, 1981; O’Brien & Myers, 1985; Tyler, Hertel, McCallum, & Ellis, 1979). All of this evidence suggests that infrequent words are more salient.

Many studies on pronoun comprehension have shown that factors that influence the saliency of an antecedent, such as sentence and discourse focus, affect the processing of pronouns (Garnham, 2001). Therefore, if frequency influences saliency, pronouns with infrequent antecedents should be easier to read than pronouns with frequent antecedents, because infrequent antecedents are more salient.

Unfortunately, results from previous studies are mixed. Pynte and Colonna (2000) investigated syntactically ambiguous sentences in which a relative clause could be attached to one of two possible noun phrases that preceded it. In their eye-movement experiment, they observed that readers initially preferred to attach the relative clause to the least frequent noun phrase, supporting the saliency account. However, establishing a syntactic relation between two phrases may be very different from establishing a coreference relation between a pronoun and an antecedent.

An experiment by Simner and Smyth (1999) suggests that antecedent frequency effects do not occur during pronominal processing. In their experiment, participants read sentences containing a pronoun that referred to an antecedent that was presented as a picture. The picture represented either a frequent or infrequent word. However, this manipulation did not have any effect on the reading times for a sentence containing the pronoun. Similarly, in a clause-by-clause self-paced reading study, Simner, Smyth, and Garnham (2003) did not find an antecedent verb frequency effect in the processing of verb phrase ellipses. Thus, the experiments by Simner and colleagues support the lemma reaccess account.

However, we think that the experiments by Simner and colleagues are not conclusive. In both experiments, they used a self-paced reading task in which reading times were measured for whole clauses containing an anaphor. We believe that this task may not be sufficiently sensitive to detect antecedent frequency effects, because such effects could be relatively small and localized. Furthermore, in both experiments, the antecedent was very salient: it was either visually present in a picture, or it was the only verb phrase in the clause preceding the anaphor. If the saliency account is correct, the saliency of the antecedent may have been too high for frequency to have any additional effect.1

---

1 Indeed, pilot studies that we conducted showed no antecedent frequency effects for pronouns when the antecedent was in focus (i.e. when it was the subject of the preceding clause).
In our experiment, we used eye-tracking methodology, which can detect small and localized effects such as lexical frequency (Rayner & Duffy, 1986; Schilling et al., 1998). Furthermore, the antecedent was always in object position, and was therefore not the main focus of the sentence (Gordon, Grosz, & Gilliom, 1993). Finally, because the lemma reaccess account predicts no antecedent frequency effect, we tried to pre-empt a possible null effect by comparing pronouns with repeated noun phrase anaphors. Rayner, Raney, and Pollatsek (1995) have shown that frequency effects persist even on second mention of a noun. If our frequency manipulation is sufficiently strong, we expect to observe a similar effect for repeated NPs in our experiment. Note that this effect may be due to either lexical access of the repeated noun or reaccess of the antecedent. According to the full reaccess hypothesis, both factors should contribute to the effect, whereas the lemma reaccess account predicts that it is caused by lexical access of the repeated noun. Finally, although the saliency account predicts that reaccess of the antecedent should be easier when it is infrequent, this effect may be overridden by the lexical access effect for the repeated noun.

2. Experiment

2.1. Method

2.1.1. Participants
Fifty-two native speakers of English from the University of Dundee participated in the experiment. All had normal vision.

2.1.2. Items
We constructed 36 materials similar to the example in Table 1.² Two factors were crossed: Frequency of the antecedent (frequent vs. infrequent) and Anaphor type (pronoun vs. repeated NP). The CELEX English database (Baayen, Piepenbrock, & Van Rijn, 1993) showed that the type frequency of the frequent antecedents (mean 98.3 per million words, range 24.9–304.6) was much higher than that of the infrequent antecedents (mean 1.2 per million, range 0.1–3.4). The length of the high and low frequent noun never differed by more than one character.

The materials consisted of two sentences. The anaphor of interest was either a possessive pronoun or a genitive noun phrase that was part of a prepositional phrase in the second sentence. Its antecedent was always the direct object of the preceding sentence. All pronouns were unambiguous; they could not refer to any other noun phrase.

2.1.3. Design
The design was a fully factorial repeated measurements Latin square design. We constructed four lists of items, with exactly one version of each item appearing in each list. Thus, each participant saw each item and each condition, but never saw more than one version of the same item. Each antecedent-noun pair (e.g. criminal/arsonist) was used in two different items. However, each noun occurred only once in each list. Each list also

² The list of materials can be obtained from the first author.
contained 60 fillers. Participants received a comprehension question for 46 of the trials. The experimental materials and fillers were placed in a single random order.

2.1.4. Procedure

Participants’ eye movements were recorded with a Fourward Technologies Dual Purkinje Generation 5.5 eye tracker which monitored the right eye (though viewing was binocular). The tracker had an angular resolution of 10' arc. The screen displayed 3.8 characters per degree of visual angle. The tracker monitored participants’ gaze location every millisecond to establish the sequence of eye fixations and their start and finish times.

Participants were instructed to read the sentences carefully in order to understand them, but to read at his or her normal rate. Bite bars and head restraints were used to minimize head movements. Before each item a calibration check was performed, and the participant was recalibrated if necessary.

2.1.5. Analyses

Trials with major tracker losses were excluded (3.0% of trials). If a fixation was shorter than 80 ms and within one character space of the previous or next fixation, it was assimilated to that fixation. We excluded all remaining fixations shorter than 80 ms and fixations longer than 800 ms.

We defined three regions for analysis purposes (Table 1). The *antecedent region* consisted of the antecedent noun. The *anaphor region* consisted of the pronoun or repeated noun plus the three character spaces to the left of it. Including fixations to the left of the region of interest is a commonly used procedure for analyzing short regions, because it is assumed that short words such as pronouns are processed during a fixation close to the left of the word when they are skipped (Ehrlich & Rayner, 1983; Garrod, Freudenthal, & Boyle, 1994). The *post-anaphor region* included the word immediately following the anaphor. However, if this word was shorter than five characters, the next word was included as well.

We analyzed three eye-tracking measures. *First-fixation time* is the first fixation in a region provided the reader has not fixated subsequent text. *First-pass time* is the sum of all fixation times starting with the reader’s first fixation inside a region until the reader’s gaze...
leaves the region (either to the left or right), again provided that the reader has not fixated subsequent text. For regions consisting of a single word, this corresponds to gaze duration (e.g. Rayner & Duffy, 1986).

Total time is the sum of all fixations in a region. Skips of a region in a particular measure were treated as missing data points.  

2.2. Results

Table 2 presents the mean reading times by condition. For each eye-tracking measure and each region, we conducted two ANOVAs, one with subjects (F₁) and one with items (F₂) as the random factor. The ANOVAs contained Anaphor type and Frequency as within-participants and within-items factors, and Participant/Item Group as a between-participants and items factor (Pollatsek & Well, 1995). Because we were only interested in the frequency effect and its interaction with anaphor type, we do not report main effects of anaphor type, which are most likely due to length differences and lexical differences between nouns and pronouns.

Table 2

<table>
<thead>
<tr>
<th>Region</th>
<th>Antecedent</th>
<th>Anaphor</th>
<th>Post-anaphor</th>
</tr>
</thead>
<tbody>
<tr>
<td>First-fixation times</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repeated NP – infrequent antecedent</td>
<td>276 (5)</td>
<td>250 (4)</td>
<td>252 (4)</td>
</tr>
<tr>
<td>Repeated NP – frequent antecedent</td>
<td>245 (3)</td>
<td>234 (3)</td>
<td>254 (4)</td>
</tr>
<tr>
<td>Pronoun – infrequent antecedent</td>
<td>284 (5)</td>
<td>245 (4)</td>
<td>233 (4)</td>
</tr>
<tr>
<td>Pronoun – frequent antecedent</td>
<td>247 (4)</td>
<td>250 (5)</td>
<td>250 (4)</td>
</tr>
<tr>
<td>First-pass reading times</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repeated NP – infrequent antecedent</td>
<td>434 (13)</td>
<td>376 (10)</td>
<td>281 (5)</td>
</tr>
<tr>
<td>Repeated NP – frequent antecedent</td>
<td>296 (6)</td>
<td>311 (7)</td>
<td>282 (5)</td>
</tr>
<tr>
<td>Pronoun – infrequent antecedent</td>
<td>437 (13)</td>
<td>281 (7)</td>
<td>264 (6)</td>
</tr>
<tr>
<td>Pronoun – frequent antecedent</td>
<td>305 (7)</td>
<td>297 (8)</td>
<td>288 (7)</td>
</tr>
<tr>
<td>Total times</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repeated NP – infrequent antecedent</td>
<td>596 (17)</td>
<td>450 (13)</td>
<td>336 (8)</td>
</tr>
<tr>
<td>Repeated NP – frequent antecedent</td>
<td>379 (10)</td>
<td>379 (9)</td>
<td>333 (8)</td>
</tr>
<tr>
<td>Pronoun – infrequent antecedent</td>
<td>599 (18)</td>
<td>362 (10)</td>
<td>339 (10)</td>
</tr>
<tr>
<td>Pronoun – frequent antecedent</td>
<td>378 (10)</td>
<td>373 (11)</td>
<td>356 (10)</td>
</tr>
</tbody>
</table>

First-fixation, first-pass, regression-path and total times are reported in ms. Standard errors are in parentheses.}

leaves the region (either to the left or right), again provided that the reader has not fixated subsequent text. For regions consisting of a single word, this corresponds to gaze duration (e.g. Rayner & Duffy, 1986). Total time is the sum of all fixations in a region. Skips of a region in a particular measure were treated as missing data points.  

3 We also analyzed first-pass regressions (regressions after a first-pass fixation) and regression-path times (sum of fixations from first entering the region until going past it). No effects of frequency or interactions with anaphor type were observed in first-pass regressions. The pattern of means in regression-path times was nearly identical to that in first-pass times, but the effect of frequency and the interaction with anaphor type was not significant in the post-anaphor region. To check whether there was a frequency effect just before the anaphor, we analyzed the region of the second sentence that preceded the anaphor region. No effects of frequency or interactions with type of anaphor reached standard levels of significance.
2.2.1. Antecedent region
First-fixation, first-pass, and total times showed that infrequent antecedents took longer to read than frequent antecedents (first-fixation: $F_1(1, 48) = 52.62$, $P < 0.01$; $F_2(1, 32) = 56.02$, $P < 0.01$; first-pass times: $F_1(1, 48) = 98.06$, $P < 0.01$; $F_2(1, 32) = 129.86$, $P < 0.01$; total times: $F_1(1, 48) = 134.95$, $P < 0.01$; $F_2(1, 32) = 110.44$, $P < 0.01$). In contrast, there was no significant difference between pronouns with frequent and infrequent antecedents (first-pass times: $F_1(1, 48) = 42$, $P > 0.05$; and total times: $F_1(1, 48) = 58$, $P > 0.05$). This latter finding supports the saliency account. No effects of frequency or interaction with anaphor type occurred in total times ($F_1 < 1.75$).

2.2.2. Anaphor region
In first-fixation times, the main effect of frequency was not significant ($F_1(1, 48) = 3.58$, $P = 0.07$; $F_2(1, 32) = 2.75$, $P = 0.11$), but there was an interaction between frequency and type of anaphor ($F_1(1, 48) = 8.30$, $P < 0.01$; $F_2(1, 32) = 8.67$, $P < 0.01$). Simple effects showed that infrequent repeated NPs took longer to read (250 ms) than frequent NPs (234 ms) ($F_1(1, 48) = 14.42$, $P < 0.01$; $F_2(1, 32) = 11.66$, $P < 0.01$). In contrast, there was no significant difference between pronouns with frequent and infrequent antecedents (250 vs. 245 ms, $F < 1$).

We observed a main effect of frequency in first-pass ($F_1(1, 48) = 9.89$, $P < 0.01$; $F_2(1, 32) = 9.14$, $P < 0.01$) and total times ($F_1(1, 48) = 9.29$, $P < 0.01$; $F_2(1, 32) = 8.95$, $P < 0.01$), but this was qualified by an interaction between frequency and anaphor type (first-pass times: $F_1(1, 48) = 21.29$, $P < 0.01$; $F_2(1, 32) = 48.45$, $P < 0.01$; total times: $F_1(1, 48) = 17.18$, $P < 0.01$; $F_2(1, 32) = 13.73$, $P < 0.01$). Infrequent repeated NPs were read slower than frequent NPs in both first-pass (376 vs. 311 ms; $F_1(1, 48) = 19.92$, $P < 0.01$; $F_2(1, 32) = 53.92$, $P < 0.01$) and total times (450 vs. 379 ms; $F_1(1, 48) = 16.55$, $P < 0.01$; $F_2(1, 32) = 21.60$, $P < 0.01$). There were no significant differences between pronouns with frequent and infrequent antecedents (first-pass: $F_1(1, 48) = 3.40$, $P = 0.07$; $F_2(1, 32) = 2.47$, $P = 0.13$; total times: $F < 1$).

2.2.3. Post-anaphor region
We observed a main effect of frequency in both first-fixation ($F_1(1, 48) = 7.86$, $P < 0.01$; $F_2(1, 32) = 8.04$, $P < 0.01$) and first-pass times ($F_1(1, 48) = 6.89$, $P < 0.01$; $F_2(1, 32) = 5.85$, $P = 0.02$), but again, this effect was qualified by an interaction between frequency and anaphor type (first-fixation: $F_1(1, 48) = 4.01$, $P = 0.05$; $F_2(1, 32) = 4.65$, $P = 0.04$; first-pass times: $F_1(1, 48) = 4.42$, $P = 0.04$; $F_2(1, 32) = 7.80$, $P < 0.01$). In contrast to the anaphor region, simple effects showed no differences between frequent and infrequent NPs (all $F < 1$), whereas reading times were faster after pronouns with infrequent than frequent antecedents (233 vs. 250 ms in first-fixation; $F_1(1, 48) = 10.33$, $P < 0.01$; $F_2(1, 32) = 10.18$, $P < 0.01$; 264 vs. 288 ms in first-pass times; $F_1(1, 48) = 8.64$, $P < 0.01$; $F_2(1, 32) = 10.29$, $P < 0.01$). This latter finding supports the saliency account. No effects of frequency or interaction with anaphor type occurred in total times ($F < 1.75$).

3. Discussion
First-fixation, first-pass, and total times for the antecedent showed that infrequent antecedents took longer to read than frequent antecedents, consistent with other
eye-movement studies on lexical frequency (Rayner & Duffy, 1986; Schilling et al., 1998). Furthermore, first-pass and total times for infrequent NP anaphors were also longer than for frequent NP anaphors. This is consistent with Rayner et al. (1995), who showed that frequency effects persisted on the second occurrence of a word. No effects occurred in the region following the NP anaphor.

More importantly, in the pronoun conditions, we observed frequency effects in first-fixation and first-pass times for the region following the pronouns. Reading times after pronouns with infrequent antecedents were faster than after pronouns with frequent antecedents. This provides evidence that lexical properties of the antecedent can affect the processing of pronouns. The data support the saliency account, which claims that infrequent antecedents are easier to access than frequent antecedents, because they are more salient. As we have suggested in Section 1, this may be because infrequent antecedents attract more attention than frequent antecedents, and as a result, are encoded more strongly. Furthermore, infrequent words require more processing effort and this may also result in better encoding.

The frequency effect for pronouns did not occur at the pronoun itself. This is not surprising, as other eye-movement experiments have also observed that the earliest effects occur after the pronoun (Ehrlich & Rayner, 1983; Garrod et al., 1994). The processing of pronouns may be slightly delayed because they require access of a discourse representation, and discourse effects may occur later than lower-level lexical effects. This explains why the frequency effect occurred early for NP anaphors, but late for pronouns. The frequency effect for NP anaphors is most likely due to lexical access of the repeated noun and therefore occurs early. In contrast, the effect after pronouns is due to reaccess of the antecedent. This explanation is consistent with the saliency account, because it claims that antecedent frequency influences the saliency of the antecedent in the discourse representation, and should therefore affect the ease with which a pronoun can refer to it.

Our results are inconsistent with the predictions of the full reaccess hypothesis, according to which reaccessing an antecedent involves the same processes as lexical access. Given that lexical access for infrequent words is slower than for frequent words, reading times for pronouns with infrequent antecedents should be slower. However, we observed the opposite pattern.

The lemma reaccess account does not explain our results either. It claims that the processing of pronouns involves reaccess of the antecedent’s lemma, but not its lexeme. Lemma information does not include frequency, and therefore retrieval of the antecedent’s lexical information should not result in a frequency effect. However, it is possible to accommodate our results into this type of account by positing two distinct stages during antecedent reaccess. During the first stage, the processor identifies the antecedent. The more salient it is, the easier it is to identify, so infrequent antecedents are reaccessed faster than frequent antecedents. During the second stage, the processor reaccesses lexical properties of the antecedent, which includes lemma information only. Thus, although Simner and Smyth’s (1999) account is currently incomplete as to how antecedent frequency affects pronoun processing, it could account for our findings by incorporating the saliency explanation for antecedent frequency effects.
Acknowledgements

This research was partly supported by British Academy grant no SG-35247, awarded to RvG, and a European Union Marie Curie Fellowship, awarded to AM. The authors are solely responsible for information communicated and the European Commission is not responsible for any views or results expressed. We would like to thank Keith Rayner, Tony Sanford and Julia Simner for discussions and encouragement, and Chuck Clifton for the use of his eye-tracking software.

References


