

## Constraints on Optional *that*: A Strong Word Form OCP Effect

Mary Ann Walter and T. Florian Jaeger  
MIT and Stanford/MIT

### Introduction\*

When used in its function as complementizer or relativizer, the English word that can often be omitted. Examples (1) and (2) illustrate that omission in complement clauses (henceforth CCs) and in non-subject-extracted relative clauses (henceforth NSRCs), respectively.

- (1) I realize [(that) John wanted to see you].
- (2) I bought Peter [<sub>NP</sub> the book [(that) you recommended \_]].

In addition to its use as complementizer or relativizer (henceforth optional that), that can also occur as a demonstrative determiner (henceforth determiner that) or pronoun (henceforth pronoun that). So, if the embedded subject of a complement or relative clause begins with determiner that, (3a), or if the subject is a pronoun that, (3b), speakers have the option to avoid uttering two adjacent thats by omitting the optional that.

- (3) a. I believe (that) that drug makes you sleepy.  
b. I believe (that) that makes you sleepy.

We employ this phenomenon to investigate the effect of the Obligatory Contour Principle, a constraint that avoids adjacent identical elements (henceforth OCP, Leben 1973), on lexical/constructional choice.

OCP effects have been investigated both in terms of what *kind of identity* is avoided and what *strategies* are used for avoidance. The kind of identity has been of interest to researchers because susceptibility to the OCP entails, and thus provides evidence for, the existence of that mental representation.<sup>1</sup> OCP-motivated analyses have been proposed for the following cases, thereby furnishing support for the array of linguistic primitives given in column two (Ross 1972; Leben 1973; McCarthy 1986; Billings and Rudin 1996; Yip 1998; Rohdenburg 1998; Richards 2001; Walter 2005):

---

\* We are grateful to Laura Staum for helpful feedback on earlier versions of this paper. This work was supported in part by an NSF Graduate Research Fellowship to the first author and by Stanford Summer Research Assistantship of the second author.

<sup>1</sup> Recent work (e.g. Frisch, Pierrehumbert, and Broe 2004) has argued that the OCP is computed with respect to similarity rather than identity (with identity being a special case of similarity). We hold that similarity at one level may still be attributed to identity at lower levels of representation which then have cumulative effects varying in strength.

(4)	a. place of articulation	articulatory gestures/features
	b. individual speech sounds	phonemes
	c. lexical tone	autosegmental (tonal) tiers
	d. affixes	morphemes
	e. determiners/wh-words	words
	f. functional heads	syntactic categories
	g. syntactic constructions	phrasal constituents

Note that these effects are not confined to the traditionally phonological sphere for which the OCP was first proposed. Rather, they seem to encompass most if not all levels of linguistic representation.<sup>2</sup>

Strategies for repairing OCP violations are similarly wide-ranging, and include deletion of one of the identical forms; form alternation of one of the identical elements (whether at the allophonic or allomorphic level); insertion of additional intervening material; and constituent reordering (whether via metathesis, syntactic movement, or some other mechanism).

One arena which has not yet been investigated, however, is the application of the OCP at the word form level. Double that sequences, as in (3), constitute a rare kind of environment in which two identical word forms can appear in immediate proximity (i.e. without intervening phonological material, including prosodic breaks). In double that sequences, the adjacent items are identical only in terms of the abstract representation of their word forms. The two adjacent thats come from different lemmas and belong to different syntactic classes – relativizer/complementizer and determiner/pronoun. This distinguishes them from word strings of type (4e). Furthermore, the phonetic realization of optional that often differs from the phonetic realization of pronoun and determiner that (Jurafsky, Bell, and Girard 2002) in a way that cannot be reduced to contextual factors (they occur in different prosodic, syntactic, and phonological environments). Jurafsky et al. found that optional that is frequently subject to vowel reduction (while pronoun/determiner that almost never is)<sup>3</sup>, and optional that is never accented (but determiner/pronoun that frequently is). They also found differences in the mean length between different uses of that (see also Berkenfield 2000). Thus double that sequences are also qualitatively different from purely phonetic OCP effects.

Finally, optional that omission provides an environment in which identical adjacent words can be avoided without resulting in ungrammaticality (and as we argue below, at least in some cases without resulting in a change of meaning). Given the optionality of relativizer/complementizer that, a repair to the OCP-violating double that sequence is readily available. For this reason, optional that is

---

<sup>2</sup> Perceptual effects of a similar type may also be observed in vision and other cognitive domains (e.g. repetition blindness, Kanwisher 1987).

<sup>3</sup> In at least one American English dialect, the two forms of that have vowels of different quality altogether, independent of reduction (Elliot Moreton, personal communication).

an optimal testing ground for word form OCP effects. If the OCP affects optional that, then that omission should be significantly more frequent before embedded subjects that begin with (or are) that, compared to other embedded subjects. We dub this the Word Form OCP Hypothesis.

Section 2 reviews factors known to influence the distribution of optional that. Section 3 presents two studies arguing that the Word Form OCP Hypothesis holds for pronoun that. Section 4 consists of three studies arguing that the OCP effect also applies before determiner that. Both sets of studies present data from syntactically parsed corpora (Penn Treebank III) and from the World Wide Web (WWW). While the WWW data is noisier, it allows us to investigate the word form OCP Hypothesis for a variety of genres and registers. Also, for some questions, there was not enough data available in the parsed corpora. Section 5 provides a summary and discusses the consequences for future research on OCP effects as well as some consequences for models of that omission.

### **Optional Complementizers and Relativizers**

The choice of optional that over the competing zero form has been attributed to at least three different sources: (a) a difference in linguistic meaning (e.g. Dor 2005; Thompson and Mulac 1991); (b) differences in style/register (Adamson 1992; Huddleston and Pullum 2002); and (c) differences in processing efficiency (e.g. Hakes, Evans, and Brannon 1976; Hawkins 2004, 2001; Ferreira and Dell 2000; Race and MacDonald 2003; Rohdenburg 1998; Jaeger and Wasow 2005). We see no reason for these different approaches to be mutually exclusive. Indeed, there is evidence that all of the above sources influence that omission. Optional that correlates with measures of processing complexity, such as the presence of speech disfluency in the embedded clause (Jaeger 2005), the grammatical weight of the embedded clause (Race and MacDonald 2003; Roland, Elman, and Ferreira in press; Jaeger, Orr, and Wasow 2005), and the predictability of the embedded clause (Wasow and Jaeger 2005; Jaeger, Orr, and Wasow 2005). This suggests that at least some of the variation in that omission is not due to difference in meaning (whether social or linguistic) associated with optional that. At the same time, register (Ferreira and Dell 2000) and the gender of speakers (Jaeger and Staum to be presented) influence optional that frequency beyond the known processing factors. Hence a meaningful test of the Word Form OCP Hypothesis can only be conducted while other factors are controlled.

The Word Form OCP Hypothesis makes predictions about the frequency of optional that before embedded subjects that begin with that compared to other types of subjects. Crucially, that omission has also been shown to correlate with the complexity of the embedded subject: the more complex the subject of a CC or NSRC, the more likely is optional that (Roland, Elman, and Ferreira in press for complementizer omission; Jaeger and Wasow 2005 for relativizer omission). So in order to have an adequate baseline for the comparison of that omission before embedded subjects with initial that, it is especially important to find another type of subject that differs minimally from that-initial subjects.

In the studies presented below, we used demonstrative this (either in its function as pronoun or in its function as determiner) as baseline. Demonstrative this is suitable for the comparison with demonstrative that since both forms are similar in meaning (deictic determiners), and have similar phonological forms (the same syllable structure, phoneme count, and onset; approximately equal duration).<sup>4</sup> Furthermore, pronoun this and that have been shown to correlate with the same degree of discourse accessibility (Gundel, Hedberg, and Zacharski 1993). This is relevant since higher accessibility of CC and NSRC subjects has been linked to the likelihood of optional that (Jaeger and Wasow 2005). Unlike the demonstrative pronouns, *determiner this* and *that* have been argued to differ in discourse accessibility (Gundel, Hedberg, and Zacharski 1993). Crucially, however, the difference biases the results *against* the Word Form OCP Hypothesis.<sup>5</sup> Some of the studies presented below include additional controls, which are discussed as part of those studies.

### **An initial test: Optional that before pronoun that**

As an initial test of the Word Form OCP Hypothesis, we compared the distribution of optional that before pronoun that and pronoun this in two studies on different corpora. The first study is based on the WWW. The second study confirms the results on the parsed Penn Treebank III corpora (Marcus et al. 1999). Even though the first study is arguably less controlled, we include it to show the pervasiveness of the OCP effect across different corpora.

The Word Form OCP Hypothesis predicts that optional that is omitted significantly more frequently if the adjacent embedded subject is the demonstrative pronoun that. As tests on the Treebank III corpora did not indicate that that omission is affected differently by the presence of an adjacent pronoun that in CCs versus NSRCs, we present the pooled results for that omission in the two clause types.

### **Study 1: World Wide Omission**

We searched the WWW using the search engines Google and AltaVista, using two different search engines since it is known that the results of web searches

---

<sup>4</sup> Both syllable/stress structure and the complexity of the onset of words influence the phonetic reduction of preceding words (Bell et al. 2003), in the present case optional that. So even if word omission is partly driven by phonetic reduction – or, more generally, even if that omission is partly driven by the phonetic/phonological environment – demonstrative this is a suitable baseline.

<sup>5</sup> The more accessible the subject of NSRC, the less likely is a relativizer (Jaeger and Wasow 2005). For example, NSRCs with a pronominal subject lack a relativizer more frequently than NSRCs with a lexical subject. This calls for caution when comparing clauses with embedded subjects of different degrees of accessibility, as arguably we do by comparing this N and that N. But as this N has been argued to refer to *more* accessible referents than that N (Gundel, Hedberg, and Zacharski 1993), there should be *more* relativizers before that N is than before this N is. The Word Form OCP Hypothesis predicts the opposite. Thus the baseline we chose should result in a conservative test of the Word Form OCP Hypothesis.

sometimes differ depending on the search engine. All searches were performed on 07/18/2005 and included only English pages located on servers in the USA.<sup>6</sup> Below we are not interested in the absolute number of matches for any search, but rather the relative ratios of the number of matches. So we use the number of *pages* containing matches to a search (which is what Google returns) as an estimate of the number of matches.

To get an estimate of the rate of optional that before pronoun that, we conducted searches for the strings that is and that that is. The rate of optional that was calculated by dividing this number by the number of hits for that is (the estimate for the number of *all* CCs and NSRCs with a pronoun that subject). The rate of optional that before pronoun that was then compared to the rate of optional that before pronoun this subjects (which was calculated parallel to the pronoun that cases).<sup>7</sup>

The inclusion of the copula is into the search strings serves three purposes. First, including the verb into the search string excludes cases of determiner that/this. Second, keeping the embedded verb constant reduces variation in that omission due to other factors (see previous section). We chose the verb is because it is highly frequent (and therefore guarantees a high number of matches for our searches). Third, we chose the singular form because it is necessary to exclude instances of that followed by a verb with plural agreement. While that and this share most of the environments in which they can be followed by a plural verb (e.g. All people doing this/that are cowards), there are environments in which only that can be followed by a plural verb: at the beginning of a CC or NSRC that can be followed by a plural verb (namely when that is optional that). Using a singular-marked verb form (here the copula is) excludes those case.

---

<sup>6</sup> While the reliability of web searches has been questioned (see Veronis 2005 and references therein), most of the known problems pertain only to one-word searches. All searches presented here matched strings of words. Keller and Lapata (2003) show that online data can be used for reliable frequency estimates that resemble the distribution of carefully balanced corpora.

<sup>7</sup> Note that this procedure rests on three crucial assumptions that we did not test. We see no reason to believe that these assumptions are problematic, but we mention them here for the sake of completeness. First, the search strings that is and this is do not only match cases we are interested in but rather *any* demonstrative subject NP (embedded in a CC/NSRC or not). Thus, if for some reason there are proportionally more false positives of this kind for this is than for that is, this could cause a false alarm (i.e. an unjustified rejection of the null hypothesis in favor of the Word Form OCP Hypothesis). Second, our search strings that that is and that this is do not distinguish between CCs and NSRCs. However, CCs have a higher base rate of that omission. So, if for some reason pronoun that subjects are proportionally more frequent in CCs (than in NSRCs) compared to pronoun this subjects, this, too, could result in a false alarm. Finally, search engines do not filter out punctuation. For example, the search string that that is also matches "... I cannot believe I missed *that*. *That is* so stupid. ...". So, if for some reason this occurs more frequently immediately preceding punctuation than that, this could result in a false alarm. All of these potential problems with the search patterns do not matter here as long as they affect the that- and this-search strings in the same way. Furthermore, Study 2 avoids the above-mentioned problems and confirms the results of Study 1.

Comparing over 260,000,000 instances of that is to over 370,000,000 instances of this is, we found that optional that was much less likely before pronoun that than before pronoun this (about 15.8 times according to Google, and about 8.9 times according to AltaVista; Fisher's Exact  $P < 0.001$ ).<sup>8</sup>

This result provides initial support for the Word Form OCP Hypothesis, which predicts avoidance of optional that before pronoun that (to avoid double that sequences). But could it be that, despite our effort to find an adequate baseline, this difference is driven by some property of pronoun that or pronoun this that has nothing to do with the OCP? The most striking difference between the two pronouns is the [+/- proximate] feature. To control for an effect of this difference, we calculated the rates of that omission before the strings these are and those are (these and those also differ in terms of the [+/- proximate] feature). We then normalized the rate of that omission before the singular demonstrative pronouns by the rate of that omission before the plural demonstrative pronouns. Even after controlling for the [+/- proximate] difference, optional that is still much less likely before pronoun that than before pronoun this (about 9.9 times according to Google, and about 5.8 times according to AltaVista). The distribution of optional that before pronoun that on the WWW thus provides support for the Word Form OCP Hypothesis.

### **Study 2: Penn Treebank III**

We recognize that there are a number of problems with web searches. So, we conducted a second study on the Penn Treebank III corpora (Marcus et al. 1999), which consist of about 800,000 parsed and part-of-speech tagged sentences from spoken language and 1.3 million sentences from written language. The syntactic annotation of the Penn Treebank III corpora is quite reliable (all parses have been hand-checked by syntactically trained annotators). In our dataset, we estimate the rate of false inclusion to be less than 1%. We used Tgrep 2 (Rohde 2001) to extract CCs and NSRCs out of the Wall Street Journal corpus (written English; henceforth WSJ), Brown Corpus (written English from a variety of genres, see Francis and Kucera 1979, henceforth BC), and Switchboard Corpus (informal spoken English from transcribed telephone conversations between two strangers on selected topics, see Godfrey, Holliman, and McDaniel 1992, henceforth SWBD). We found 17,813 CCs and 6,576 NSRCs in the Penn Treebank (to reduce noise in the dataset only CCs immediately adjacent to their embedding verb and NSRCs with either no relativizer or optional that were included).

---

<sup>8</sup> Additional searches were performed including strings in which the copula forms a contraction with the pronominal subject (e.g. that is → that's). The results were qualitatively the same, though contraction was observed to be more frequent following zero (90%) than optional that (80%). This complies with the generalization that syntactic deletion is often accompanied by phonological reduction (e.g. Dressler 1972).

Since we found only 21 NSRCs with pronoun that subjects, we only report the results for CCs below. Averaged across all three corpora, CCs were introduced by optional that in about 28% of all cases.

We found a total of 830 CCs beginning with a pronoun this or that subject in the Penn Treebank corpora. The distribution of optional that (% OPT) is reported in Table 1. The last row of Table 1 gives the significance level of Fisher's Exact tests (two-sided) for each of the three corpora and across the three corpora (Total). All three corpora display the same trend that was observed in Study 1: optional that is less frequent before pronoun that than before pronoun this, on average 3.1 times less frequent. The OCP effect is highly significant for the pooled results (see the last column of Table 1). The difference also reaches significance for the WSJ and SWBD (BC lacks enough data to reach significance).

**Table 1** Rate of optional that (% OPT) before pronoun that and this in CCs

Subject	WSJ		BC		SWBD		Total	
	% OPT	Total	% OPT	Total	% OPT	Total	% OPT	Total
Pronoun <u>this</u>	26%	39	58%	19	18%	50	28%	108
Pronoun <u>that</u>	2%	41	33%	6	9%	675	9%	722
Fisher's Exact	<b>p&lt;0.01</b>		<b>n.s.</b>		<b>p&lt;0.05</b>		<b>p&lt;0.001</b>	

As an additional control, we extracted data on that omission before pronoun these and those subjects, but did not find enough cases for statistical testing (N=21).

### Intermediate Summary

As predicted by the Word Form OCP Hypothesis, the results from Study 1 and Study 2 show significantly less optional that immediately preceding pronoun that than one would expect a priori. Shortcomings of the web searches presented in Study 1 were addressed in the second study. Taken together, the two studies show that double that sequences are avoided across a wide range of genres and styles/registers (which at least the WWW data averages over). Furthermore, the effect does not seem to be limited to written language (cf. the SWBD results). We will return to this point in the general discussion.

### Optional that before determiner that

The Word Form OCP Hypothesis also predicts an OCP effect for CC and NSRC subjects with determiner that, as in (3a), repeated below. This prediction is tested in the next three studies.

- (3) a. I believe (that) that drug makes you sleepy.

Study 3 is based on web searches. Study 4 uses the Linguist's Search Engine (Elkiss and Resnik 2004) to conduct syntactic searches on corpora gathered on the WWW. Study 5 uses the Penn Treebank III corpora.

### Study 3: World Wide Omission

Parallel to Study 1, we used Google to compare the rate of optional that before determiner that and determiner this (AltaVista searches were also conducted but are not reported here since they do not differ from the Google results in any relevant way). All searches were performed on 07/17/2005 (only English pages located on servers in the USA). This time we were interested in matches to the searches (that) that/this ... is, where ... is the rest of a subject NP introduced by the demonstrative determiner. To reduce false inclusions (as well as unwanted variation due to different types of subject NPs), we used a list of singular nouns (see Table 2 below) and searched for the strings that N is (this N is) and that that N is (that this N is). For each N, we calculated and compared the rate of optional that before this N is and that N is (just as done, *mutatis mutandis*, in Study 1).

The inclusion of the copula is into the search string has been discussed in Study 1. Here including the copula has an additional purpose. It excludes hits in which N is not the complete subject, e.g. phrasal compounds, as in (5a,b).

- (5) a. I think that *peer to peer* file sharing is wrong.  
b. I think that *boy meets girl plots* are the best kind.

Only Ns that do not lend themselves to use as bare nouns were used. This was done since the polysemy of that combined with singular noun subjects that can be used with and without a determiner (the latter would be a bare noun) can result in ambiguity that lasts for several words after the embedded verb or even does not get resolved at all. Consider the following case, where that can either be optional that followed by a bare noun (generic) NP beer or the beginning of the NP that beer referring to a specific, deictically identified brand of beer:

- (6) I told you that beer from my hometown is bad.

Since speakers may avoid such ambiguities, we exclude them from the current study. We return to the issue of ambiguity avoidance in the general discussion.

To see whether the OCP effect holds across different types of embedded subjects, we tested three nouns of each of the following seven types: monosyllabic animate nouns (e.g. girl), monosyllabic inanimate nouns (e.g. bed), disyllabic animate nouns with initial stress (e.g. husband), disyllabic inanimate nouns with initial stress (e.g. picture), nouns with three to four syllables and initial stress (e.g. capitol), nouns with three to four syllables and stress on the second syllable (e.g. infection), nouns with the primary stress on a later syllable (e.g. explanation).

Table 2 summarizes the mean ratio of optional that before that vs. this for each of the seven types of nouns. The Word Form OCP Hypothesis predicts values larger than 1, which indicate that optional that was less frequent before that N is than before this N is for that noun.

**Table 2** Optional that rates before this N is vs. that N is

Type of noun			Total number of hits		(%OPT before <u>this N</u> ) / (%OPT before <u>that N</u> )
Syll.	Stressed	Animate	<u>this N is</u>	<u>that N is</u>	
1	1 <sup>st</sup>	Y	881,142	913,805	7.8
1	1 <sup>st</sup>	N	14,846,405	147,494	3.9
2	1 <sup>st</sup>	Y	208,670	90,261	8.7
2	1 <sup>st</sup>	N	1,964,900	191,060	1.5
3+	1 <sup>st</sup>	-	44,250	44,435	67.6
3+	2 <sup>nd</sup>	-	2,984,294	80,323	10.8
3+	Later	-	236,902	61,420	14.7
<b>OVERALL</b>			<b>21,166,563</b>	<b>1,528,798</b>	<b>3.4</b>

An analysis of variance (ANOVA) with the nouns (not the types of nouns) as random factor and determiner type (that vs. this) as only fixed effect confirmed that optional that is significantly less frequent before determiner that ( $F(1,20)=23.7$ ,  $p<0.001$ ). The control comparison of optional that before these N vs. those N did not reach significance ( $F(1,19)=1.1$ ,  $p>0.3$ ; one noun search did not yield enough hits): optional that seems to be equally frequent before determiner these and determiner those.

To conclude, the web data supports the Word Form OCP Hypothesis (the lack of an effect for the control comparison means that the effect cannot be due to the [+/-proximate] feature). Optional that is less frequent before determiner that than before determiner this, despite the fact that omitting that before determiner that introduces (temporary) ambiguity. Although we found quite a bit of variation in the strength of the OCP effect (cf. last column of Table 2), this seems to be due to differences between the nouns, not between the types of nouns (we found that the optional that rates differ as much or more *within* each group of nouns as the differ *between* the different types). We thus tentatively conclude that the OCP effect for determiner that is independent of properties of the subject noun.

The high number of matches returned by web searches enabled us to hold several factors known to influence the distribution of optional that stable (the embedded verb and the complexity of the embedded subject). But web searches contain considerable noise. So we replicated the results on more controlled corpora, the LSE Web Collection and the Penn Treebank III corpora. These studies are discussed next.

#### **Study 4: Linguist's Search Engine and World Wide Trees**

The LSE Web Collection consists of 3.5 million parsed and part-of-speech tagged sentences from 175,000 documents gathered on the web. We used the Linguist's Search Engine (Elkiss and Resnik 2004) to extract 546 CCs and 31 NSRCs that start with a determiner that or this.

For CCs, we found that optional that before determiner that (total N=72) is about 2.9 times less frequent than before determiner this (total N=474). This difference is significant (Fisher’s Exact  $p < 0.001$ ). The 31 NSRCs are not enough for a meaningful statistical test, but numerically we observed the same tendency as for CCs. Optional that in NSRCs beginning with determiner that is about 2.7 times less frequent than in NSRCs beginning with determiner this. Only 11 CCs and 2 NSRCs starting with determiner these or those were found, which made a comparison to this vs. that impossible. Nevertheless, data from the LSE Web Collection confirm the trend observed in Study 3 and provide further support for the Word Form OCP Hypothesis.

### Study 5: Penn Treebank III

The purpose of Study 5 was to confirm the results observed on the WWW by a more controlled study on the Penn Treebank III corpora. We used the same databases as in Study 2. We only found 11 instances of NSRC subject beginning with that or this (of which only one was preceded by optional that), too few for any meaningful investigation. In the 17,813 CCs of the Penn Treebank corpora, we found a total of 93 CCs with a subject beginning with determiner that or this.

Table 3 summarizes the results. Overall optional that is about 2.7 times less frequent before CC subjects that begin with determiner that compared to subjects that begin with determiner this (two-sided Fisher’s Exact  $p < 0.02$ ). The result also reaches significance for the WSJ, but not for the SWBD and BC.

**Table 3** Rate of optional that (% OPT) before determiner that and this in CCs

Subject	WSJ		BC		SWBD		Total	
	% OPT	Total	% OPT	Total	% OPT	Total	% OPT	Total
Pronoun <u>this</u>	22%	18	50%	6	35%	17	32%	41
Pronoun <u>that</u>	3%	32	50%	2	22%	18	12%	52
Fisher’s Exact	<b>p=0.05</b>		<b>n.s.</b>		<b>n.s.</b>		<b>p&lt;0.02</b>	

The results are consistent with the Word Form OCP Hypothesis. Only two of the corpora contain enough examples for a meaningful test (WSJ, SWBD). While both of these corpora display the same trend, only one of the tests reaches significance. Furthermore, we found that the overall difference of optional that before determiner that [-proximate] vs. this [+proximate] is only slightly stronger than the difference before determiner those [-proximate] vs. these [+proximate]: optional that is about 2.2 times less frequent before determiner those (25%) than before determiner these (50%) and this difference is marginally significant (Fisher’s Exact  $p < 0.06$ , N=54). While the results for that vs. this resemble the results for those vs. these *averaged across corpora*, the effect is observed in different corpora. The only somewhat significant ( $p=0.1$ ) effect for those vs. these comes from the Switchboard corpus. Given the small number of occurrences, it is not possible to test whether optional that is disfavored before determiner that even after the effect of the [+/-proximate] feature is controlled for.

### Intermediate Summary

All three studies (using five different corpora consisting of different genres, registers, and modes) show significantly less optional that before determiner that than before determiner this. Study 3 suggests that this is not due to the [+/- proximate] difference between this and that. Study 4 did not yield enough data to confirm this. Study 5 shows a slightly larger asymmetry between determiner that vs. this than between determiner those vs. these, but the difference is too subtle to draw a strong conclusion. We tentatively conclude that there is at least some evidence suggesting that the avoidance of optional that before determiner that is not due to the [+/- proximate] feature.

### General discussion

We have presented evidence that the OCP has an effect on the distribution of optional that even after other factors known to correlate with optional that are controlled. Before pronoun that and before determiner that, optional that is significantly less frequent than expected. The OCP effect is strong (double that sequences are on average between 2 to 4 times less frequent than expected given the baseline). Furthermore, our studies take into account only one avoidance mechanism (dropping optional that). The potential use of other ways of avoiding OCP violations (e.g. constituent reordering such as topicalization in the embedded clause; the choice of which instead of optional that for nonhuman head nouns in NSRCs; insertion of intervening material such as fillers or editing terms like you know or uh/um, or adverbials such as to the best of my knowledge; etc.) means that our tests provide a conservative estimate of the word form OCP effect. In sum, the results support the Word Form OCP Hypothesis.

The word form OCP effect reveals itself as a statistical bias rather than a categorical generalization. Double that sequences remain grammatical, though dispreferred. In this respect it patterns with OCP phenomena like those documented by Frisch et al. (2004), in which a place-of-articulation OCP effect holds with varying strength based on both similarity of the segments in question, and distance of those segments from each other. Coetzee (to appear) documents a similar and likewise gradient effect in perception (of place, for English). If these phenomena form part of the grammar, then, they provide evidence for theories of grammar that can account for gradient intra-speaker variation (e.g. Boersma and Hayes 2001; Anttila 2002). Alternatively, the effect observed here could be due to processing preferences. We leave this question to future research.

A related question for future research is whether adjacent identical word forms are hard to *produce* and/or hard to *comprehend*. Answering this question could help to distinguish production-facilitation accounts (Ferreira and Dell 2000; Race and MacDonald 2003; Jaeger and Wasow 2005) and comprehension-facilitation accounts of optional that (Hawkins 1994, 2004; Rohdenburg 1996; Temperley 2003). Interestingly, the distribution of optional that before embedded subject NPs that can occur equally well with or without a determiner (cf. Study 3)

offers an opportunity for future research to test comprehension-facilitation accounts in which ambiguity avoidance is a strong factor (Hawkins 1994, 2004; Temperley 2003). We briefly elaborate on this.

As mentioned in Study 3, embedded subject head nouns that can occur either with or without a determiner (in both cases being in the singular) can cause ambiguity that does not get resolved by the singular marking on the noun or the singular agreement of the verb. Recall (6), repeated below, where that can either be optional that followed by a bare noun (generic) NP beer or the beginning of the NP that beer referring to a specific, deictically identified brand of beer:

(6) I told you that beer from my hometown is bad.

Crucially, speakers can avoid the ambiguity in (6). If the that in (6) is an optional that, then dropping it prevents ambiguity. If the that is a determiner, then inserting optional that before it prevents ambiguity. Ambiguity avoidance accounts therefore predict *more* double that sequences before ambiguous subject nouns (such as beer) that are introduced by determiner that; they predict *fewer* double that sequences if the ambiguous subject noun is used without a determiner.

The polysemy of that can also introduce temporary ambiguity if the subject head noun cannot be used as a bare noun. This is the case because it is the number marking on the head noun that forces disambiguation (singular marking on a noun that cannot be used as a bare noun forces a preceding that to be read as demonstrative that). So, if determiner that is followed by several adjectives, which do not necessarily disambiguate between different lemmas of that, the ambiguity persists:

(7) I told you that *big fat and vicious* ...  
a. ... raccoons are dangerous.  
b. ... raccoon is dangerous.

Comprehension-facilitation accounts in which ambiguity avoidance is a strong factor (Hawkins 1994, 2004; Temperley 2003) predict that disambiguating double that sequences become *more* likely the longer the ambiguity introduced by a single that would last.<sup>9</sup> We leave these predictions to future research since our current data set does not contain enough CCs or NSRCs with potentially ambiguous subject head nouns or complex subject NPs as in (7).

---

<sup>9</sup> Even if prosodic and contextual cues as well as phonetic differences between the different lemmas of that (see Section 1) help addressees to distinguish the different lemmas of that, double that sequences may still be considered the more explicit marking. Comprehension-facilitation accounts that assume that speakers reduce processing costs for addressees in environments of increased processing complexity (e.g. Hawkins 2004; Rohdenburg 1996) would therefore nevertheless predict *more* optional that before determiner that than before determiner this.

We now address a possible objection to our conclusion. One could object to our conclusion that the results argue for the *Word Form* OCP Hypothesis. Orthographic identity holds in the case of two thats as well as word form identity, and thus is an alternative source of an OCP effect. But while most of the results we have presented are based on written language, we have included results from spoken language, for which the observed effect is smaller but still significant. We thus tentatively conclude that the observed OCP effect is at least in part driven by word form identity. We leave it to future research to determine how much of the effect in written corpora is due to orthographic identity. Ongoing research on non-orthographically identical English homophone strings teases apart the relative contributions of these dimensions of identity (Walter in progress).

Finally, the Word Form Hypothesis makes a prediction that we have not tested. Utterances that contain OCP-violating double that sequences must have some property that otherwise tends to be associated with higher likelihood of optional that. That is, CCs and NSRCs with double that should be more complex than otherwise comparable clauses (i.e. CCs and NSRCs with this-initial subjects). Further research is necessary to verify this point and determine what kind(s) of complexity can outweigh the dispreference for double that.

## Conclusion

Despite its potential functional role in early disambiguation, speakers prefer to omit optional that if a sequence of identical lexical items would result. This preference appears to be an active if gradient effect, with strong consequences for online linguistic production. Finally, the existence of yet another domain of application for the OCP calls into question the proposal (e.g. of Boersma 1998) that what seems to be a uniform identity-avoidance phenomenon is actually a constellation of unrelated constraints. We conclude that the OCP applies to the (phonological) word form as well as to other grammatical categories discussed above.

## References

- Adamson, H. Douglas. 1992. Social and Processing Constraints on Relative Clauses. *American Speech* 67 (2).
- Anttila, Arto. 2002. Variation in phonological theory. In *The handbook of language variation and change*, edited by J. K. Chambers, P. Trudgill and N. Schilling-Estes. Malden, Mass.: Blackwell Publishers.
- Bell, Alan, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand, Michelle Gregory, and Daniel Gildea. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustic Society of America* 113 (2):1001-1024.
- Berkenfield, Catie. 2000. The role of syntactic constructions and frequency in the realization of English that. Master thesis, University of New Mexico, Albuquerque, NM.
- Billings, Loren, and Catherine Rudin. 1996. Optimality and superiority: A new approach to overt multiple-wh ordering. In *Proceedings of the annual workshop on formal approaches to Slavic linguistics (the College Park meeting)*, edited by J. Toman. Ann Arbor: Michigan Slavic Publications.
- Boersma, Paul. 1998. *Functional phonology*. The Hague: Holland Academic Graphics.

- Boersma, Paul, and Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32:45-86.
- Coetzee, Andries. to appear. The OCP in the perception of English. In *Prosodies. Selected Papers from the Phonetics and Phonology in Iberia Conference, 2003*, edited by S. Frota, M. Vigarío and M. J. Freitas. New York: Mouton de Gruyter.
- Dor, Daniel. 2005. Toward a semantic account of that-deletion in English. *Linguistics* 43 (2):345–382.
- Dressler, Wolfgang U. 1972. Approaches to Fast Speech Rules. In *Phonologica*, edited by W. U. Dressler and M. W. Fink.
- Elkiss, Aaron, and Philip Resnik. 2004. The Linguist's Search Engine User Guide. University of Maryland.
- Ferreira, V.S., and G.S. Dell. 2000. Effect of Ambiguity and Lexical Availability on Syntactic and Lexical Production. *Cognitive Psychology* 40:296-340.
- Francis, W. N., and H. Kucera. 1979. Brown Corpus Manual. Brown University.
- Frisch, S. A., J. A. Pierrehumbert, and M. B. Broe. 2004. Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* 22:179-228.
- Godfrey, J., E. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. Paper read at ICASSP-92.
- Gundel, Jeanette K., Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language* 69 (2):274–307.
- Hakes, D. T., J. S. Evans, and L. L. Brannon. 1976. Understanding sentences with relative clauses. *Memory & Cognition* 4 (3):283-290.
- Hawkins, John A. 1994. *A Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press.
- . 2001. Why are categories adjacent? *Journal of Linguistics* 37:1-34.
- . 2004. *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.
- Huddleston, Rodney, and Geoffrey Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge UP.
- Jaeger, T. Florian. 2005. Optional *that* indicates production difficulties: Evidence from disfluencies. Paper read at Disfluency in Spontaneous Speech Workshop (DiSS'05), September 09-12, 2005, at Aix-en-Provence.
- Jaeger, T. Florian, David Orr, and Thomas Wasow. 2005. Comparing Frequency-based and Complexity-based Accounts (of Relativizer Omission). The 18th Annual CUNY Sentence Processing Conference, March 31st - April 2nd, 2005, Tuscon, AZ.
- Jaeger, T. Florian, and Laura Staum. to be presented. That-Omission Beyond Processing: Stylistic and Social Effects. NWAV.
- Jaeger, T. Florian, and Thomas Wasow. 2005. Processing as a Source of Accessibility Effects on Variation. Paper read at Berkeley Linguistic Society.
- . 2005. Production Complexity Driven Variation: The case of relativizer distribution in non-subject-extracted relative clauses. The 18th Annual CUNY Sentence Processing Conference, March 31st - April 2nd, 2005, Tuscon, AZ.
- Jurafsky, Daniel, Alan Bell, and Cynthia Girand. 2002. The Role of the Lemma in Form Variation. In *Laboratory Phonology VII*, edited by C. Gussenhoven and N. Warner. Berlin/New York: Mouton de Gruyter.
- Kanwisher. 1987. Repetition blindness: Type recognition without token individuation. *Cognition* 27:117-143.
- Keller, Frank, and M. Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics* 29 (3):459-484.
- Leben, Will. 1973. *Suprasegmental Phonology*, Linguistics, MIT, Cambridge.
- Marcus, Mitchell P., Beatrice Santorini, Mary Ann Marcinkiewicz, and A. Taylor. *Treebank III*. Linguistic Data Consortium 1999 [cited].
- McCarthy, J. 1986. OCP Effects: Gemination and antigemination. *Linguistic Inquiry* 17:207-263.

- Race, David S., and Maryellen C. MacDonald. 2003. The use of "that" in the production and comprehension of object relative clauses. Paper read at 26th Annual Meeting of the Cognitive Science Society.
- Richards, N. 2001. A distinctness condition on linearization. MIT.
- Rohde, D. 2001. Tgrep2 Manual. Brain & Cognitive Science Department, MIT.
- Rohdenburg, Günter. 1996. Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7 (2):149-182.
- . 1998. Clausal complementation and cognitive complexity in English. Paper read at Anglistentag, at Erfurt, Germany.
- . 1998. Cognitive complexity and *horror aequi* as factors determining the use of interrogative clause linkers in English. In *Determinants of grammatical variation in English*, edited by G. Rohdenburg and B. Mondorf. Berlin: Mouton de Gruyter.
- Roland, Douglas, Jeffrey L. Elman, and Victor S. Ferreira. in press. Why is that? Structural prediction and ambiguity resolution in a very large corpus of English sentences. *Cognition*.
- Ross, J. R. 1972. Doubl-ing. *Linguistic Inquiry* 3:61-86.
- Temperley, David. 2003. Ambiguity avoidance in English relative clauses. *Language* 79 (3):464-484.
- Thompson, Sandra A., and Anthony Mulac. 1991. The discourse conditions for the use of complementizer that in conversational English. *Journal of Pragmatics* 15:237- 251.
- Veronis, Jean. 2005. Web: Google's missing pages: mystery solved? In *Technologies du Langage*. <http://aixtal.blogspot.com/2005/02/web-googles-missing-pages-mystery.html>.
- Walter, Mary Ann. 2005. The distinctness condition in Semitic syntax. MIT.
- . in progress. Homophone sequences in English. MIT.
- Wasow, Thomas, and T. Florian Jaeger. 2005. Lexical Variation in Relativizer Frequency. Expecting the unexpected: Exceptions in Grammar Workshop at the 27th Annual Meeting of the German Linguistic Association.
- Yip, M. 1998. Identity avoidance in phonology and morphology. In *Morphology and its relation to phonology and syntax*, edited by S. LaPointe, D. Brentari and P. Farrell. Stanford, CA: CSLI Publications.