Is *that* a lexical OCP effect?*

Mary Ann Walter
Massachusetts Institute of Technology

notes – golston style deciding between multiple syn outputs.
not categorical still. do surfacing that that's display other properties militating for that-inclusion? (ie material between head noun and clause)
probabilistic constraints
email flo for correls, also avoiding ambig ref
other id lex seqs. like to too vs to also – lex subst?
nb often surface phon not id due to v reduction, wanna contraction

claim that
people definitely change their outputs to avoid ambiguity. Again that's
what I figured, but if we're going to talk as if it's an established fact
we should probably cite something. I don't know of any such study, though...do you?

(Biber et al. 1999; Bolinger 1972; Temperley 2003) are claimed to claim this ;-). It's a good idea to check them. This is for relativizers. For complementizers this has been claimed by may people (I believe Hawkins 1994, in press gives references; it should also be in more or less any psycholinguistic paper on complement clause processing, maybe Ferreira & Dell, 2000 contains a summary? I attach that one).

a phonological expl, not phonetic – not case of extreme reduction, probably, see results of Mara's data where significant reduction was not found phrase-internally. include results of elicitation study in which (hopefully) significant differences appear in duration, V quality, etc. do these great differences result from need to lessen ambiguity prosodically since ocp prevents it from being done lexically?

1. Introduction

* My thanks to Florian Jaeger for useful discussion and his invaluable help with corpora.

A growing body of literature details the production and processing of the English relativizer *that*. The relativizer is typically described as optional in English. This optionality, in conjunction with the polyvalence of the word, leads to ambiguities in English sentences – phrases beginning with the word often may be interpreted either as relative clauses, or as noun phrases in which *that* is interpreted as either a pronoun or a determiner.

In this study I investigate the appearance of *that* as a relativizer when its inclusion results, or would result, in a consecutive sequence of it and its homophone(s). A theory of production relying on altruism on the part of the speaker predicts its greater inclusion in such contexts, since this resolves the ambiguity described above. However, on theories of least articulatory effort or avoidance of adjacent identical elements (in the vein of the phonological Obligatory Contour Principle), the relativizer is predicted to appear less often than otherwise expected when it precedes a *that*-initial clause. The second prediction is borne out, as the relativizer is used significantly less frequently than expected in such contexts ($p<.01$).

In Section 2 below I detail properties of the word *that* in English – its optionality as relativizer, multiple uses and the ambiguity these introduce; the conditions on its presence as a relativizer; and its potential for consecutive tokens. In 3 I give the results of a set of investigations into these potential double-that sequences, using corpora and online data. In 4 I discuss the implications of these results for the issues mentioned above.

2. The relativizer *that* and where it appears

2.1 Relativizer drop and multiple uses result in ambiguity

The relativizer *that* can be optionally dropped in English in sentences like (1a) below (though not when the clause is fronted as in 1b; examples taken from Tabor *et al*. 1997):

1. *Relativizers*
a. The lawyer insisted that cheap hotels were clean and comfortable.
b. That cheap hotels were clean and comfortable surprised us.

The same orthographic and phonological form, however, may also be used as a determiner, as shown by the near-minimal pairs in (2).

2. *Determiners*
a. The lawyer insisted that cheap hotel was clean and comfortable.
b. That cheap hotel was clean and comfortable to our surprise.

Note that pronominal variants, though not mentioned by Tabor and colleagues, are also possible:

3. *Pronouns*
a. The lawyer insisted that was clean and comfortable.
b. That was clean and comfortable to our surprise.

In what follows, the pronominal use of that will be considered together with the determiner use.

This array of options introduces considerable ambiguity for the listener, who cannot be sure which sense of *that* is appropriate until the last segment of the third word of the clauses given above. This is on the assumption that plural is regularly marked – for subject nouns for which it is not, the listener must wait until the following verb for disambiguation to occur.

Experiments show that listeners have clear preferences in how they resolve such ambiguities. Tabor *et al*. (1997) observe that when that occurs sentence-initially, as in examples (1b) and (2b), subjects tend to interpret it as a determiner. When it follows the main-clause verb, however, as in (1a) and (2a), they tend to treat it as a relativizer. This effect is attributed to subjects' use of information about differences in the frequencies of the two interpretations in the respective syntactic structures. Gibson (submitted) maintains that the difference is due to the interaction of context-free frequencies of the item, and top-down expectations about syntactic structure (independent of the frequencies of *that* tokens in it). For our purposes, it is enough to note that competing interpretations exist and exact some cost in processing for the listener.

2.2 Conditions governing relativizer drop

In addition to such work on the processing of *that* relativizers, considerable information is also available on its production and the conditions governing it. In studies conducted using the Switchboard corpus (Godfrey *et al*. 1992), which includes approximately one million words, Jaeger *et al*. (2004a) identify over four thousand non-subject extracted relative clauses. Each one is potentially initiated by a *that* relativizer. In fact, a relativizer appears roughly 50% of the time (37.1% headed by that, 15.4% by a wh relative pronoun, and 47.5% without a relativizer). This suggests that relativizer drop is a purely optional process, with no governing conditions.

However, Jaeger and colleagues go on to show that presence vs absence of relativizing *that* is in fact much more predictable when certain factors are taken into account. These factors include the complexity of dependency domains, ease of activation/retrieval of discourse referents, and semantic/pragmatic factors that militate for or against the presence of a relative clause (2004a). To summarize, the presence of intervening material between the head noun and relative clause results in the use of a *that*-relativizer 90% of the time. Use of relativizing *that* also increases significantly along with the length of the relative clause's VP. The authors conclude that syntactic complexity strongly influences the frequency of relativizer drop. This conclusion is supported by the observation that disfluencies (indicating production difficulties) correlate inversely with relativizer drop.

Several other factors lead to *less* frequent use of the relativizer. Relative clauses in which the subject is a pronoun are headed by a relativizing *that* significantly *less* often than those with full-noun subjects, with person also exerting an effect. An accessibility hierarchy of the sort proposed by Warren and Gibson (2002) seems to be in operation. A head noun's inclusion of an exclusive or superlative strongly prefers, or even requires, a subsequent modifier like a relative clause. In that case the relativizer also appears significantly less often. Finally, less contentful nouns are also more likely to be modified

by a relative clause, but less likely for that clause to be introduced by a relativizer. By hypothesis, greater predictability of a relative clause lessens the need for it to be unambiguously signaled by an overt relativizer.[1]

2.3 Double-that utterances

Leaving aside the implications of these findings for specific theories of complexity/processing difficulty, it is now clear that relativizer drop is subject to a number of conditioning factors. Each factor considered up to this point may be due to either production or comprehension motivations. That is, the greater use of relativizers in comparatively complex utterances may be due to the speaker's greater need for time to plan the utterances. Alternatively, it could be due to the speaker's knowledge that the utterance will be difficult for the listener to comprehend, and inclusion of an overt relativizer makes comprehension somewhat easier. The relative importance of speaker laziness vs altruism cannot be gauged using these conditions.

In what follows I consider yet another one, which has not been previously considered and may be able to address this distinction – the possibility of adjacent word sequences. While the examples in 1-3 above break down each potential use of *that* individually, it is also possible to conflate them in sentences such as the following:

4. *Doubles*
a. That that (cheap hotel) was clean and comfortable surprised us.
b. The lawyer insisted that that (cheap hotel) was clean and comfortable.

Such utterances contain sequences of orthographically and phonologically identical lexical items, a rare occurrence in English. When other such sequences do occur, they are typically due to a desire for emphasis or contrast, or are due to stacking of auxiliaries. In the latter case, the items are also separated by a strong prosodic boundary in pronunciation. Examples of such utterances are given in (5) below.

5. *Non-that doubles*
a. She has beautiful green, green eyes.
b. Her eyes are *green* green, not blue-green.
c. What the problem is, is that…
d. They don't do that, but what they *do* do is…

That sequences provide the only example of such a sequence without any clear syntactic, semantic, or pragmatic function for the repetition. While such factors may influence the likelihood of relativizer inclusion, as described above, it always remains possible to drop it without changing the meaning of the utterance. Thus that-sequences are a rare and

[1] This is compatible with cross-linguistic patterns of obligatory clause-marking in a way – for example, Arabic requires a relativizer only for definite head nouns. However, it is surprising that definites, which are more likely to have a relative-clause modifier and for which the clause is therefore more predictable, are the class to require a relativizer here. This is contrary to the account given by Jaeger and colleagues, where the relativizer is more likely to be dropped when the modifier is predicted. However, a more extensive cross-linguistic survey would be required for any conclusions to be drawn.

possibly unique test case for potential effects of the OCP on the lexical level, as well as production versus comprehension motivations for relativizer drop.

3. Results

To address these issues, investigations were conducted using both corpora of English and online sources. They are discussed in turn below.

3.1 Corpus data

A search was first conducted on three corpora of English, annotated by the Penn Treebank III project: Switchboard, the Wall Street Journal corpus of written English, and the Brown corpus (Godfrey et al. 1992, Mitchell et al. 1993, Frances and Kucera 1982). Thus large samples of both spoken and written English were considered. The search identified non-subject extracted relative clauses beginning with the item *that* (in its determiner/pronominal use). Clauses beginning with *this* were also identified, for purposes of comparison. Due to its semantic and phonological similarity to *that*, the item is an appropriately minimal comparison.

Results are tabulated in Table 1, which includes both the full set of sentences, and results after those clauses introduced by a wh-word relativizer were discarded.

	that-initial	*this*-initial	total
all RCs	15	10	25
wh relativizer	8	6	14
no relativizer	6	4	10
that relativizer	1	0	1

Table 1: RCs from annotated corpora, by initial word and relativizer type.

As may be seen, *that*-initial relative clauses outnumber those beginning with *this* in all counts. However, the numbers involved are too small to be the basis for any firm conclusions. Moreover, the set contained only one utterance in which a *that*-relativizer was used.[2] Thus insufficient information is available to reveal anything about the patterning of relativizer drop in this context, much less its interaction with the many other factors influencing it that are discussed above.

3.2 Online data

A second source of information comes from online searches using the Google search engine. This technique offers much less information about context than the use of annotated corpora. Comparisons between the effect of potential identical sequences and the other influences on relativizer drop, therefore, unfortunately remain impossible.[3] However, the greater numbers involved make it a powerful tool. In what follows, I describe the results of a number of such searches.

---

[2] The set of corpus utterances, minus those with wh-word relativizers, is given in Appendix 1.

[3] Other methodological drawbacks of using Google are the inclusion of multiple instances of the same tokens to an unknown degree, and the nontransparency of how it rounds off the number of hits.

As a first step, let us establish the relative frequencies of *that* and *this* in isolation. Raw counts of Google hits are given in Table 2. By this measure, instances of the two words appear comparable in frequency. However, recall that *that* has two primary uses, whereas *this* is limited to being a determiner/pronoun. Thus Table 2 also provides a normalized value for the frequency of determiner/pronoun *that* for a better comparison with *this*, using the relative frequency of the uses of *that* given by Gibson (submitted).[4]

	that	this
raw count	1,160,000,000	1,230,000,000
normalized count	262,160,000	1,230,000,000

Table 2: Frequency counts of *that/this* in isolation.

As Table 2 shows, *that* is used considerably less frequently than *this* when limited to its comparable syntactic role (only 21% as often).[5]

In the next search, hits for *that that* sequences are compared directly with those for *that this* sequences. Results are shown in Table 3.

	that that	that this
raw count	4,700,000	16,800,000

Table 3: Frequency counts of *that* + determiner/pronoun sequences.

All else being equal, the relative frequencies established above of *that* and *this* as determiner/pronouns would lead us to expect *that that* sequences to occur 21% as often as *that this* sequences. In fact, such sequences occur 28% as often.

With only these two data points, it is difficult to know if the 7% difference is significant or not. Thus another set of searches was performed. First, 10 NP pairs of the form *that X* and *this X* were searched for, where X is the same highly frequent singular noun. Singular nouns were chosen in order to single out determiner/pronoun uses of the preceding *that*. For the same reason, nouns were selected that do not lend themselves to use as generics, which might be construed as stand-alone RC subjects with *that* functioning as a relativizer.[6] Table 4 gives means of the search results for such pairs as

[4] Gibson finds that in the Brown corpus, 77.5% of that tokens are used as relativizers, 11.1% as determiners, and 11.5% as demonstrative pronouns. Here I collapse the last two categories.

[5] This is somewhat surprising given the edge in its clause-initial frequency as determiner/pronoun, though again those numbers were too small for reliability.

[6] For example, *time* in the following two examples:

6.
a. I think that time at the beach was lots of fun. (that-determiner)
b. I think that time slips away too quickly. (that-complementizer)

Even for the non-generic singular nouns selected, such relativizer uses may occasionally appear, as in examples like the following:

7.

well as hits in which they are preceded by (relativizing) *that*.[7] Results broken down by item are available in Appendix 2.

Noun	that N	this N	%	that that N	that this N	%
means	930,200	3,553,700	54	7,706	153,964	16

Table 4: Mean frequency counts of determiner-noun sequences by determiner type and relativizer use.

As we expect based on the frequency observations above, in each case the sequences containing *this* outnumber those containing the determiner *that*.[8] A percentage was then calculated of how often *that*-determiner sequences occurred relative to sequences with *this*. These too are provided in Table 4. A paired-samples two-tailed t-test performed on these percentages indicates a significant difference (p=.003)

As mentioned in footnote 7, however, at least one potential confound arises for our interpretation of the sequences discussed above. Therefore, a second set of searches was undertaken with exactly the same items used above, in which each was also followed by the copula *is*. This ensured that no compound constructions were being wrongly identified in which the single *that* tokens were actually being used as relativizers. Table 5 contains the results from these searches. Item counts are provided in Appendix 3.

Noun	that N	this N	%	that that N	that this N	%
means	60,122	1,104,020	30	1,729	37,303	15

Table 5: Mean frequency counts of determiner-noun sequences by determiner type and relativizer use, when followed by copula.

Though the numbers involved are smaller, and the percentage difference is less pronounced, the same general pattern appears. Here too the effect is significant (p=.008).

Even for this measure, the potential problem remains that the sequences with an overt relativizer are double-counted – they are included in the counts of the determiner-N sequences. These forms in no case exceed 4.5% of the determiner-N totals, so their effect cannot be great – I am assuming that it is does not significantly affect the proportions observed. To verify this, however, the proportions were recalculated after subtracting the number of that-determiner-N hits from the number of determiner-N hits. The effect remains highly significant (p=.009).

4. Discussion and conclusions

---

a. I think that person to person file sharing is wrong.
b. I think that boy meets girl plots are the best kind.

However, such constructions are relatively rare, and in addition are usually signaled by orthographic hyphens between the words, which our search terms exclude. Therefore, I assume that they do not contribute significantly to the frequency counts.

[7] Note that mean percentages are the average of all the percentages by item, *not* the percentage of the mean numbers of hits, given to the left.

[8] In fact there was one exception, in the relativizer-less NP using *dog*.

Absolute counts of double-*that* sequences compared to *that-this* sequences indicate in greater-than-expected appearance of the optional relativizer when followed by a non-relativizing token of *that*. This seems to provide support for an altruistic motivation for relativizer retention. However, the significance of this measure cannot be verified.

When such the frequency of such sequences is measured in conjunction with multiple specific nouns following them, a different pattern emerges. If there is no effect of a determiner/pronoun *that* following a relativizing *that*, we expect roughly the same proportion of that-that-N to that-this-N sequences as of that-N to this-N sequences. However, this is not the case. The former proportion is significantly smaller than the latter proportion for both of the measures calculated. That is, the relativizer appeared significantly less often when followed by its homophone.[9] Despite its potential functional role in early disambiguation, then, speakers prefer to omit the relativizer if a sequence of identical lexical items would result. Least effort/OCP appears to outweigh altruism.

[9] Note that the effect may be even stronger than observed here if other avoidance mechanisms are used in addition to relativizer drop.

References

Bresnan, J., Carletta J., Crouch R., Nissim M., Steedman M., Wasow T. & Zaenen, A. (2002). *Paraphrase analysis for improved generation*, LINK project, HRCR Edinburgh-CLSI Stanford.

Ferreira, V. S. & Dell, G.S. (2000) Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, 40, 296-340.

Francis, W. N. & Kucera, H. (1982). *Frequency analysis of english usage: Lexicon and grammar.* Boston: Houghton Mifflin.

Gibson, E. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition* 68, 1-76.

Gibson, E. (submitted). The interaction of top-down and bottom-up statistics in syntactic ambiguity resolution. MIT manuscript, Department of Brain and Cognitive Sciences.

Godfrey, J., Holliman, E. & McDaniel, J. (1992). SWITCH-BOARD: Telephone Speech Corpus for Research and Development. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 517-520. San Francisco, USA.

Hawkins, J.A. 1994. *A Performance Theory of Order and Constituency.* Cambridge University Press, Cambridge.

Hock, H.

Jaeger, T.F., Wasow T. & Orr D. (2004a). Performance & grammar: A production study of relativizer drop. Linglunch presentation, Department of Linguistics and Philosophy, MIT.

-----. (2004b). The origin of frequency effects? Semantic biases in relative clause production. Psychosemantics reading group presentation, Department of Linguistics and Philosophy, MIT.

Kayne, R.

McCarthy, J. 1986. OCP Effects: Gemination and Antigemination. *Linguistic Inquiry 17*, 207-263.

Mitchell, P.M., Santorini, B., & Marcinkiewicz, M.A. (1993). Building a large annotated corpus of english: The Penn treebank. *Computational Linguistics*, 19(2).

Mitchell, P.M., Santorini B., Marcinkiewicz M.A., & Taylor, A. (1999). Treebank-3. Linguistic Data Consortium, University of Pennsylvania.

Race, D.S. & M.C. MacDonald. 2003. The use of "that" in the production and comprehension of object relative clauses. *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*.

Richards, N. 2001. A distinctness condition on linearization. MIT ms.

Rohde, D. 2001. Tgrep2 Manual. At: http://tedlab.mit.edu/~dr/TGrep2/index.html

Ross,

Russell, K.

Tabor, W., Juliano, C. & Tanenhaus, M.K. (1997). Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes*, 12, 211-272.

Walter, M. A. 2002. Syntactic OCP effects: Akkadian construct state and relative clauses. MIT ms.

Warren, T. & Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition. 85,* 79-112.

Yip, M. 1988. The Obligatory Contour Principle and Phonological Rules: A Loss of Identity. *Linguistic Inquiry 19,* 65-100.

Zaenen, A., Carletta J., Garretson G., Bresnan J., Koontz-Garboden A., Nikitina T., O'Connor M.C., Wasow T. (2004). Animacy Encoding in English: why and how. *Proceedings of the 2004 ACL Workshop on Discourse Annotation D*. Byron and B. Webber, eds. Barcelona.

Appendix 1: Corpus utterances

1. *That*

a. But a takeover battle opens up the possibility of a bidding war with all ***that*** implies (wsj)

b. It reduces lawsuits from disgruntled employees and ex-employees with all ***that*** means for reduced legal costs and better public relations (wsj)

c. It increases employee commitment to the company with all ***that*** means for efficiency and quality control (wsj)

d. Erasing the differences still dividing Europe and the vast international reordering ***that*** implies (wsj)

e. and probably the only way ***that*** can happen is for um governments to realize that they have to pay if companies do nt (swbd)

f. and all ***that*** does is lawlessness (swbd)

g. There s no way ***that*** could work (swbd)

h. and there s no way ***that that*** can be done (swbd)

2. *This*

a. The reason ***this*** is getting so much visibility is that some started shipping and announced early availability (wsj)

b. All ***this*** has really established is MCA and the Bronfmans have agreed on a price at which they can be bought out (wsj)

c. all ***this*** is going to do is give you a little spending money while you re there (swbd)

d. now this needs to be do- this is the time ***this*** needs to be done (swbd)

e. Well that s the way ***this*** was (swbd)

f. that s the way ***this*** is done (swbd)

Appendix 2: Google results by item, no copula.

Noun	that N	this N	%	that that N	that this N	%
thing	685,000	2,220,000	31	6,560	45,400	14
place	998,000	3,110,000	32	3,510	50,100	7
person	2,620,000	3,530,000	74	53,500	128,000	42
boy	177,000	311,000	57	1,830	11,200	16
girl	698,000	920,000	76	2,720	22,200	12
cat	108,000	208,000	52	835	3,370	25
dog	376,000	288,000	131	1,650	7,180	23
book	1,270,000	9,290,000	14	3,050	1,040,000	1
article	1,060,000	13,600,000	8	2,190	224,000	1
computer	1,310,000	2,060,000	64	1,210	8,190	15

Appendix 3: Google results by item, with copula.

Noun	that N is	this N is	%	that that N is	that this N is	
thing	60,400	231,000	26	730	12,000	6
place	62,500	438,000	14	520	15,900	3
person	339,000	417,000	81	12,900	31,800	41
boy	15,200	30,500	50	245	1,680	15
girl	28,100	106,000	27	950	3,880	24
cat	7,980	29,000	28	118	740	16
dog	11,100	24,100	46	315	983	32
book	41,700	5,330,000	1	1,030	285,000	1
article	25,600	4,390,000	1	340	19,300	2
computer	9,640	44,600	22	142	1,750	8