THE MORPHOLEXICAL NATURE OF ENGLISH to-CONTRACTION

GEOFFREY K. PULLUM

University of California, Santa Cruz

The forms represented orthographically as (wanna), (hafta), (gonna), (gotta), (oughta), (usta), and (sposta) have standardly been analyzed as involving a syntactic rule or cliticization operation called to-contraction. Occasionally it has been suggested that the forms in question have been ‘lexicalized’, i.e., wanna and hafta are synchronically distinct lexemes from want and have. I argue here that neither approach is correct. The syntactic accounts are wrong to assume that the relation between wanna and want to must be syntactic, and the lexicalization accounts are wrong to assume that there is no synchronic relation: the link is one of derivational morphology. A morpholexical rule suffixes /tu/ ~ /ta/ to the base lexemes to form derived lexemes such as wanna. These to-derivatives are headed morphological structures, as described by Stump 1994. They inflect on their heads, not their edges; they are synonymous with their bases but have different subcategorization and more colloquial style associations. Various morphological and phonological idiosyncrasies indicate that the derived lexemes are morphologically compound, but their sharing of the lexical idiosyncrasies of the base lexemes show that they contain those bases as heads. All the syntactic phenomena that have been claimed to be relevant to the debate over to-contraction fall into place under the assumptions advocated here, and some new insights emerge, particularly with regard to the ‘liberal dialects’ where the pronunciation written (wanna) has a wider distribution than in most American dialects.*

1. INTRODUCTION. The publication of George Lakoff’s article ‘Global rules’ in this journal over a quarter of a century ago (1970) initiated a voluminous literature on the phenomenon known as to-contraction. The syntactic problem that sparked the debate was first noted by Lawrence Horn (see Lakoff 1970: 632): 1a is ambiguous between an ambitious reading entailing that I want to succeed Teddy and an altruistic reading entailing that I want Teddy to succeed, but most speakers find that 1b is not; it has only the ambitious reading.

(1) a. Teddy is the man I want to succeed.
   b. Teddy is the man I wanna succeed.

I will refer to these data as the Teddy facts. They can be illustrated via an actual ungrammaticality as opposed to a loss of ambiguity: 2a is grammatical and

* I had more than a little help from my friends with this article. Ivan Sag reinterested me in the topic by suggesting in early 1994 that the merits of a lexical account of to-contraction might have been underestimated. Arnold Zwicky contributed a core insight (about the relevance of Greg Stump’s work) in August 1995, and we copresented an ancestor of this paper at the 1996 LSA annual meeting under the title ‘Enter a new headed morphological structure: Exit a mythical contraction rule.’ Paul Postal, my coauthor on four earlier papers about the data treated here, provided searching critical commentary on another ancestor circulated in February 1996 as Linguistics Research Center paper no. LRC-96-01. Philip Miller and Eric Potsdam also gave written comments. None of these scholars should be assumed to endorse what I say; responsibility for all errors is mine. The errors would have been worse and more numerous without the feedback I received from audiences at Stanford University, the Santa Cruz and Berkeley campuses of the University of California, the Australian National University, and the specific helpful comments of Joyce Tang Boyland, Joan Bresnan, Sandra Chung, Chuck Fillmore, Jorge Hankamer, Rodney Huddleston, Bill Ladusaw, George Lakoff, Terry Langendoen, Jim McCloskey, and Tom Wasow. I gratefully acknowledge the specific contributions of Language editor Mark Aronof, associate editor Brian Joseph, and their two referees.
unambiguous, entailing that I want Teddy to succeed Bobby, and in classical TG terms would have had a deep structure roughly like 2b; but 2c is ungrammatical for speakers who have the constraint illustrated by the Teddy facts.

(2) a. Teddy is the man I want to succeed Bobby.
    b. $\{s\text{Teddy is }[\text{NP the man,}][s \text{ I want } [s \text{ [NP which man,}]] \text{ to succeed Bobby}]]
    c. *Teddy is the man I wanna succeed Bobby.

Most speakers of American English agree on the Teddy facts (though not all; see §7.8). Lakoff saw them as evidence for a global syntactic constraint on a phonological rule. A rule replacing want to by wanna\(^1\) was taken to apply in the phonological component but to be blocked if a noun phrase had intervened between want and the to at a premovement stage of the derivation, as in 2b. Much debate ensued.\(^2\) Yet for all its vigor, this polemical literature never succeeded in capturing the full range of phenomena.\(^3\) I will outline an account of

\(^1\) In presenting the earlier literature, where claims about the phonetic, phonological, morphological, and lexical status of cited forms are not made explicit. I follow my sources and simply write wanna (or whatever the form may be) in italics for the word form under discussion. When presenting my own analysis below I attempt to distinguish more carefully between, for example, the lexeme want, the word form wants, the spelling (wants), the phonological representation /ˈwants/, the phonetic representations such as [ˈwànts], and so on.

\(^2\) Baker (1971:180) and Baker and Brame (1971:59) pointed out that Lakoff’s constraint mistakenly blocked 1b as well, since then-current assumptions postulated a deep structure subject between want and to that was erased by a transformation called equi-NP deletion. Bresnan (1971) reanalyzed the facts nonglobally, employing an encliticization rule applying in the syntactic cycle. Selkirk (1972) posited a somewhat different encliticization account. Chomsky (1976) and many others since proposed that syntactic traces were responsible for blocking the contraction rule. Frantz (1978) and Postal and Pullum (1978) independently proposed that contraction is syntactically restricted to equi-subject clause union contexts. Bouchard (1982), Chomsky (1982), Pesetsky (1982), and subsequently others argued a very similar position, that contraction to wanna was licit only where the relation of government held between want and to. Postal and Pullum (1982) summarized most of the literature up to approximately that point. Disputes continued in several forums, but the editor of *Linguistic Inquiry* declared a moratorium on discussion of the issue in 1986. Despite the fact that essentially all proponents of the trace-theoretic analysis seem to have moved to some variant of the government approach, the trace-theoretic account continued to be treated as a basic result in generative grammar, considered uncontroversial enough to be used in pedagogical works as an exercise for students (Napoli 1993:457, Honda and O’Neil 1993).

\(^3\) Barss 1995 is the most recently published transformationalist analysis, and is not an improvement. Barss exposes a number of difficulties inherent in squaring the details of current transformational theories with an account of the blocking of contraction, and arrives at a solution involving a number of unsupported stipulations. He proposes to ‘abandon the assumption that control complements are uniformly CP, and instead consider an alternative representation. . . . in which want takes a bare IP complement’ (p. 689), solely in order to allow want to govern to and thus enable the application of a contraction rule requiring such a government relation. He fails to deal with (or even mention) the existence of the liberal dialects allowing ‘contraction’ across the trace of a wh-phrase yielding sentences like Who do you wanna die? Barss claims that given his raft of assumptions, ‘the impossibility of To-[Contraction] when the embedded subject is wh-moved . . . follows’—that is, follows from Universal Grammar. But the liberal dialects show that this is not what is wanted; the liberal dialects do not respect this UG-enforced blocking of contraction. Later in this paper (§7.8) I will offer the beginnings of a satisfying account of the properties of the liberal dialects.
wanna and similar verbs that offers a more insightful account of these forms' syntactic distribution than any of the analyses previously proposed in the literature, and makes sense of a significant set of largely overlooked morphological and phonological properties of these forms. The novel feature of the account offered here is that it involves derivational morphology.

2. THE THERAPY VERBS. The lexeme want is not the only one that shows phonetic amalgamations with infinitival to. Lakoff (1970: 632n) cited facts similar to the Teddy facts (noted by Charles Fillmore) involving the verb have:

(3) a. There are some puzzles I have to work on.
   b. There are some puzzles I hafta work on.

While 3a can mean either ‘I have some puzzles to work on’ or ‘There are some puzzles I must work on’, 3b clearly has only the latter reading.

There are just seven lexemes in English that are relevant in this connection. They are listed in 4.

(4) a. want, the basic desiderative verb of English; as in I want to leave; colloquial combination with to [ˈwaʃə] spelled (wanna). In colloquial British English it can be approximately synonymous with ‘ought’ (London working-class British You wanna watch yourself, mate = working-class American ‘You betta watch it, buddy’). Where necessary I will distinguish these as DESIDERATIVE WANT and OBLIGATIVE WANT.

b. Prospective go, the future-oriented temporal aspect verb expressing what Jespersen (1949 IV.14.2(3)) calls the ‘prospective present’; always complement to progressive be, as in I'm going to run; colloquial combination with to spelled (gonna), corresponding to many pronunciations, including [ˈɡənə], [ˈɡənə], [ɣnə], and in more casual speech [ɡə], [ɡə], or even [ŋə] as in [amŋə] ‘I'm gonna’.

c. Habitual used, the past-oriented temporal aspect verb expressing habitual or repeated past actions or a persisting past state; as in I used to run; colloquial combination with to [ˈjustə] sometimes spelled (usta).

d. have, expressing necessity or obligation; as in I have to run; colloquial combination with to [ˈhæftə] spelled (hafta).

e. got, also expressing necessity or obligation, used with optional perfect have, as in I've got to run; colloquial combination with to [ˈɡarə] spelled (gotta).

f. ought, morphologically a modal but taking a to infinitive; as in I ought to run; colloquial combination with to [ˈəɾə] or [ˈəɾə] spelled (oughta).

g. supposed, expressing social expectation, always in the passive; as in I'm supposed to run; colloquial combination with to [sə'postə], sometimes spelled (supposta), or [ˈspostə], sometimes spelled (sposta).
These verbs do not correspond to any traditionally recognized morphological or syntactic class. I will refer to them here as the THERAPY VERBS.4

Emonds 1977 claims that other predicates undergo to-contraction, writing nonce spellings like needa, certain 'a, and happen 'a to indicate purported contracted pronunciations of need to, certain to, and happen to, etc., and a few others (e.g. Barss 1995) have also taken such a view. But it is an error. Pronunciations like ['nida] for need to or ['hæpərə] for happen to occur only in rapid or very casual speech, reflect no influence of syntax, and are determined by automatic phonology (thus they would apply similarly to rhyming forms provided the speech style was the same). It is an important fact about the data treated here that they do not represent a phenomenon linked to rapid, casual, slovenly, or uneducated speech. Certain fast speech rules of English are relevant to understanding the etiology of the synchronic phonology, but that phonology is now lexeme-specific (see §6 for further discussion).

The case of the negative auxiliary verb forms don’t, won’t, can’t, and so forth is closely analogous (see Zwicky & Pullum 1983 for an analysis). Even in the slowest and most careful American English speech, the negative auxiliaries are not just acceptable but effectively obligatory; the periphrastic substitutes do not, etc., sound forced and unnatural in speech. By contrast, consider the pronunciation of why don’t you sometimes written as (whychna). A slow and careful delivery of Why don’t you believe me? cannot begin with whychna, but would certainly include don’t (the alternative, Why do you not believe me?, would be unusual, even stilted). And in the same way, an utterance like You are just gonna hafta accept it can be delivered slowly and distinctly in educated standard American English.

3. A WRONG TURN: THE LEXICALIZATION ANALYSIS. It has been proposed a number of times that forms such as wanna and hafta are lexical items in their own right, synchronically unrelated to want and have.5 Pullum and Postal (1979: 696) dismiss such proposals on grounds of their morphological consequences. Precisely where the alleged base form wanna would demonstrate its status as an independent verb if it took suffixes (as it would if *She wannas go were grammatical, for instance), it fails to provide such a demonstration. Yet contra Chomsky and Lasnik (1978: 271, n. 4), it is not a modal, for the modals are grammatical in the 3sg present with no -s suffix (compare She will go with *She wanna go). And contra Sag & Fodor 1994, it is not comparable to the exceptional lexeme BEWARE, which is limited entirely to the base form, i.e. the infinitive, the imperative, and ‘subjunctive’ that-clauses.6

4 I take it that the reader will not find this nonce term obscure. There are surely few of us who have never experienced a conflict between their wannas and their haftas.

5 This suggestion was made for desiderative wanna in such works as Brame 1976 p. 142 and Wasow 1991; for the British obligatory wanna in Chomsky & Lasnik 1978:271, 4; for hafta by Rotenberg 1978:47-8; for all the therapy verbs by Schmerling 1983 and Inkelas & Zec 1993:249–50; and it has been endorsed in various other published and unpublished works, the most recent being Sag & Fodor 1994.

6 That is, it is not just the third singular *John bewares [of] the dog that is ungrammatical, but also unaffixed present tense forms as in *We always beware of the dog; the discussion by Fodor 1972 fails to make this clear.
THE MORPHOLEXICAL NATURE OF ENGLISH to-CONTRACTION

The quasi-serial verb go, banned from all inflected contexts (*Every day I go fetch the paper. *I have often gone fetch(ed) the paper; see Pullum 1990), might appear to offer a better analogy. However, forms like wanna ['wānts] in He wanna go pose a problem. If wanna does not occur in the third person singular, then wanna must be a contracted pronunciation of wants to, and will share all its readings. But in fact most or all speakers whose dialects exhibit the Teddy facts have the grammaticality contrast in 5, which the defective wanna analysis counterpredicts.

(5) a. Teddy is the man Mike wants [wānts] to send.
b. Teddy is the man Mike wants [wānts] to go.
c. Teddy is the man Mike wanna [wāsta] send.
d. *Teddy is the man Mike wanna [wāsta] go.

Even clearer evidence is provided by necessitative have. If an inflectionally defective lexeme hafta with no inflected forms is postulated, we have no explanation for the lack of ambiguity in 6 unless we also postulate, implausibly, a second, completely synonymous, defective lexeme hasta, this one possessing only the 3sg present form.

(6) There are some puzzles she hasta [hæsta] work on.

Sag and Fodor refer to hasta as an ‘exceptional’ 3sg inflected form of hafta. But this is not just exceptional; it is utterly unprecedented for person inflection in a Germanic language to involve replacement of a stem-internal consonant. And, most suspiciously, the alleged consonant replacement yields exactly the form that we would expect if hafta were not a separate lexical item, but just an occurrence of have with regressive voicing assimilation on the final consonant when in construction with infinitival to. It also has the participle that treating it as have + to would lead us to expect, as seen in I resent having to clean up after them—unless we say that hafta is defective there too and gets replaced by have + to in those contexts. I conclude that taking wanna, hafta, and kin to be new verbs synchronically unrelated to want, have, etc. is completely implausible. In §4 I offer an alternative.

4. The derivational rule of to-suffixation. What is right about the lexicalization analysis, I claim, is that in modern colloquial American English, wanna and want are distinct lexemes. What is wrong about it is the claim that they are synchronically unrelated. I propose that they are related via derivational morphology.

There is a suggestion of this idea in at least one place in the prior literature. Brame (1981: 286n) suggests that we ‘take wanna to be a lexical item itself’ but then adds: ‘such an analysis does not preclude a theory which permits lexical fusion of hit and ing and of want and to to yield hitting and wanna as derived lexical words.’ In others words, wanna and hafta might be derived from want and have, respectively, by a rule of word formation in the sense of Aronoff 1976. There would be no syntactic operation of to-contraction, but there would be a derivational operation (involving the phonological operation of to-suffixation) among the word-formation rules. Central to developing this
idea explicitly (as Brame did not do) is the theory of headed morphological structures, to which I now turn.7

4.1. HEAD MARKING. Inflectional affixes sometimes seem to attach within a word, closer to the root than a derivational affix also present in the word, which is unexpected given traditional views of the relation between derivation and inflection. When we form duckling from duck and -ling we expect the plural to be ducklings, not *ducksling. This is not always the way things work, however. For example, Anderson (1992: 205ff) notes that when the Icelandic middle verb [[kall]st] (kallast) ‘be called’ is inflected with the preterite suffix /að/ and the first person plural subject agreement suffix /um/, the result is kölluðumst ‘we were called’, with the inflections closer to the root than the middle-forming derivational element -st. He suggests that the inflections are attaching to the head of kallast.

This idea, that morphological operations can apply to heads rather than affect whole words in ways shown by affix addition at their edges, can be traced back to influences of categorial grammar and extended Montague frameworks in the early 1980s.8 Stump (1994) develops the idea further, discussing several relevant cases. He takes the Sanskrit form nyapatat ‘he was flying down’ to be a headed structure, with root patat ‘fly’, derivational preverb ni- ‘down’ (acting rather like the particles of English verb-particle constructions), stem ni-patat ‘fly down’, and inflectional preterite augment a- prefixed to the head of nipatat, namely patat. Likewise, he claims that in Southern Barasano (spoken in Colombia) the derivational operation of suffixing -aka forms headed morphological structures with the semantics of diminutives; from the lexeme wi ‘house’ we can form the derived lexeme wiaka ‘little house’. The plural inflection is -ri, and since -aka diminutives are headed, the plural of wiaka is not *[[wi]aka]ri but [[[wi]ri]aka] ‘little houses’.

Stump develops a theory of this morphological construction type. He argues that where a form inflects on its head, all inflectional forms in the paradigm do likewise; that all other forms arising via the same derivation or compounding rule will also exhibit head-marking inflection; and that preservation of suppletive forms in the paradigm of a derivative is always an indication of head marking. Consider the lexeme undergo, which he claims is an example in English of a headed morphological structure. It has a stem go and a derivational prefix under-. Under- derives verbs from verbs with a change in subcategorization: go is a strictly intransitive verb but undergo is strictly transitive. The past tense form underwent, with the same suppletive past found in the head lexeme go, indicates under Stump’s theory that head marking is in play. Thus Stump’s theory predicts that head marking will occur throughout the paradigm of undergo. This is confirmed by other suppletive forms such as undergone.

---

7 This important idea is due to Arnold Zwicky.

8 Hoeksema 1985 uses head operations in a categorial grammar approach to morphology. Hoeksema’s work has links to categorial-influenced syntactic work of the time such as Pollard 1984, and earlier work like Bach 1979. It influences later developments, e.g. Aronoff 1988 and McCarthy & Prince 1990 (see pp. 225ff).
Stump also predicts that every other lexeme derived by means of the preverb *under- will also exhibit head marking, and indeed we find lexemes such as undertake, with suppletive head-marking forms undertook, undertaken (not *undertaked). Notice that under-, like similar prefixes such as over-, applies to only a short and rather unpredictable list of verb stems: compare undergo/*overgo, undertake/overtake (with semantic difference), underwrite/overwrite (the latter only with the literal meaning ‘write over’), undercover/*overgird, underpin/*overpin, *underride/overcome, *underrun/overrun, *underride/over- ride, *undersee/oversee, *underwhelm/overwhelm (the former now occurring jocularly).

I propose that the therapy verbs take a derivational postverbal element we can represent as *to which derives a headed morphological structure of the same sort as those just discussed. But where under- is a prefix, *to is a suffix, so we see the head inflection word-internally in forms like haste.

4.2. TO-DERIVATION. The morphological rule I propose, which I will refer to as MLR_{to}, applies to the seven therapy verb stems to derive synonymous lexemes of category V{[−AUX]} that are subcategorized for a bare infinitive complement and are associated with colloquial rather than formal style. Phonologically, MLR_{to}(X) = X{−tu}/ for the therapy verbs, and otherwise is not defined. Semantically, the operation is vacuous: MLR_{to}(X) denotes exactly what X denotes. In the case of WANT, for example, want'(to'(B))(α) ⇔ wanna'(B)(α)—and notice in addition that to'(B) ⇔ (B), for infinitival to is itself semantically vacuous. It is not uncommon in English for morphological rules to derive lexemes synonymous with the inputs; note the synonymy of pairs like cyclic/cyclical, comic/comical, and pairs like gold/golden, drunk/drunken, and many other such cases. Note also the contrast between treat me nice (informal) and treat me nicely (formal), where zero-derived and -ly-derived adverbs have the same semantics but different style associations.9

To-derivatives differ in subcategorization from their bases, just as the subcategorization of undergo differs from that of go. The therapy verbs take to-infinitival VP (or PRO-subject) complements, but their derivatives via MLR_{to} are subcategorized for ‘bare infinitive’ or base-form complements, like various other verbs in English (most modals; supportive do; causative verbs like make and let; perception verbs like see and hear; and the verb help, which becomes relevant in §7.3 below).

5. MORPHOLOGICAL STRUCTURE OF INFLECTED FORMS OF THE THERAPY VERBS. The theory set out in Stump 1994 predicts the right morphological structure for inflected forms of to- derivatives of therapy verbs. Take prospective go as an example. It is constrained to occur only as the head verb of the complement to the progressive auxiliary be, and thus always appears with the

---

9 It would not be correct to regard treat me nice as having an adjective substituted for an adverb. Treat takes a manner adverb complement. In certain dialects and styles a zero-derived adverb is permitted in complement function, but only there; even speakers who accept treat me nice do not accept *I was nice treated, because zero-derived adverbs are not permitted as premodifiers. See Zwicky 1995 for more on the nature of this distinction.
inflection that progressive be demands, which I will call the gerund-participle. Stump’s theory of head marking predicts correctly how the gerund-participle of a to-derivative will be formed. If M is a morpholexical rule that derives lexemes from lexemes, and PF_{[σ]} is some paradigm function (a rule specifying what the shape is for the form of the paradigm associated with the feature bundle σ), then Stump’s universal HEAD APPLICATION PRINCIPLE (1994: 263) says that if two lexemes Y and Z are related by M so that Y = M(Z), then PF_{[τ]}(Y) = M(PF_{[σ]}(Z)) if this is defined. M(PF_{[σ]}(Z)) is referred to as ‘the H-application of PF_{[σ]} to Y’. Let PF_{[ing]} be the morpholexical function that returns the gerund-participle form of a verb in English.\(^{10}\) The lexeme realized orthographically as gonna is the to-derivative of go, i.e. MLR\(_\text{to}(go)\). This lexeme is a Y such that Y = M(Z) where M = MLR\(_\text{to}\) and Z = go. Thus according to Stump’s head application principle, its -ing form will be MLR\(_\text{to}(PF_{[ing]}(go))\) if that is defined. PF_{[ing]}(go) is defined; it is the concatenation of Base(go) and -ing, i.e. going. So the form we seek is:

\[
(7) \quad \text{PF}_{[\text{ing}]}(Y) = \\
\text{MLR}_{\text{to}}(\text{PF}_{[\text{ing}]}(\text{go})) = \\
\text{MLR}_{\text{to}}(\text{Base}(\text{go})/\text{η}/) = \\
\text{MLR}_{\text{to}}(/\text{go}/\eta/) = \\
/\text{go}/\eta/\text{tu}/
\]

That is, under Stump’s theory, to obtain the ing-form of the to-derivative of go, which is derived by a lexeme-to-lexeme rule of to-derivation, compute the H-application of the to-derivation rule to it, i.e. the to-derivative of the ing-form of go, i.e. the result of suffixing -to to going, notated here as /gonτ/ (though it undergoes further lexeme-specific phonological reductions; see §6).

In a similar way, the lexeme have has a to-derivative lexeme, which we can call hafta, derived by MLR\(_\text{to}\). In the third person singular, the usual third singular -z-suffixation rule applies, and because that operation is defined for have, it applies to the head of hafta, so we get have-s-to (/hæv/z/τ/) rather than *have-to-s. There is a special allomorphic irregularity in the paradigm of have: the stem-final /v/ disappears in the third person singular, yielding [hæz] rather than [hævz] from [hæv]/z/. Stump’s theory incorporates the claim that purely morphological feature specifications do not percolate up from the head of a headed morphological structure to the whole structure (see Stump 1994: 249ff). Thus it can only be because of head marking (if wildly coincidental irregularity in two verbs is ruled out) that haste appears rather than *havesta [hævsta].

Stump’s theory also incorporates a principle, the PARADIGM UNIFORMITY GEN-

\(^{10}\) The statement of PF_{[ing]} is very simple: PF_{[ing]}(X) = Base(X)/η/, where Base(X) is the phonological shape for the base or bare infinitive form of X. Modals do not have gerund-participles because they do not have base forms for PF_{[ing]} to build on, and there are one or two radically defective verbs such as beware, but otherwise all verbs, however irregular, form gerund-participles in this way.
eralization, which says that structures that exhibit head marking do so categorically, throughout their paradigm of inflected forms (254). This entails that the hafta derivative must exhibit the same property of borrowing its forms from have at all other points in its paradigm. And sure enough, the past tense of hafta is had-to (often pronounced [ˈhætə], with regressive voicing assimilation affecting the [d] of had), not *haved-to.\textsuperscript{11}

6. The special phonology of the therapy verbs. All of the seven therapy verbs exhibit at least some special word-internal phonology, readily described in terms of idiosyncratic features of the forms of a particular lexeme but highly anomalous under the assumption of a regular syntactic contraction process.

6.1. Gonna. The form written as gonna has various pronunciations in different dialects: [ˈgonə], [ˈgɔnə], [gənə], [ɡə], [ɡo], [ŋo], etc. (They are generally not accented, since the main accent in the clause will regularly be on the main lexical verb, not the temporal aspect verb expressing the prospective present.) These pronunciations are not available for rhyming sequences such as sewing to or throwing to. It is not even available for the present participle of the motion verb go when followed by to: I’m going to hope has the alternate form I’m gonna hope, but I’m going to Hope (President Clinton’s boyhood home in Arkansas) has no gonna alternate. This can be directly confirmed from any corpus of written representations of spoken English: the spelling (gonna) will always be immediately followed by a verb.\textsuperscript{12}

The special phonology relating [ɡəntu] to [ɡənə] is not inherently surprising: nasal assimilation (or just the alveolarization seen in goin’), flapping or loss of [t] after the nasal, and vowel reduction yield the most common pronunciation. But neither is it part of the regular, automatic phonology of English; it is lexeme-particular morphophonology.

6.2. Hafta. It was stated without argument in the previous section that the devoicing of the /v/ of have and the /z/ of has seen in the phonetic forms [ˈhæftsə] and [ˈhæsta] is due to an allomorphy rule conditioned by the suffix -to. The evidence that automatic phonological rules are not responsible is straightforward enough: there is no devoicing in words like dovetail or Aztec. Lexeme-specific rules dropping consonants from clusters before suffixes, on the other hand, are known elsewhere in English morphology (note the loss of [t] in soften but not in caftan, in moisten but not in piston, and so on, due to

\textsuperscript{11} Joseph (1992:137) reports that there are speakers who say [ʰæftu] for having to. This will be discussed shortly, in \$6.

\textsuperscript{12} I tested this claim on the Wall Street Journal corpus of about 40 million words (CD ROM no. LDC93T1, Linguistic Data Consortium, University of Pennsylvania. 1993, directory WSJ). Out of 97 sentences containing uses of gonna (always in quoted speech), every one was immediately followed by a verb. Interestingly, there is also one sentence containing a prescriptivist grumble about ‘casual misspellings’ such as gonna for going to, are for our and February and February. But gonna is not a misspelling, and it is not casual, but fully conventional, as shown by the fact that it is never casually substituted for occurrences of going to in which go is a verb of motion.
lexical facts about the lexemes soft, moist, etc.). Hafia exhibits similar lexeme-specific morphophonological alternations.\(^{13}\)

As always, nonautomatic morphophonemic alternations give rise to opportunities for reanalysis, and there is an opportunity with Hafia. The devoicing rule conditioned by the voiceless initial consonant of to yields a stem that could be misanalyzed as underlying /hæf/. If this were taken to be the stem of a verb taking a to-marked infinitival complement, a paradigm like 8 might be expected.

(8) Expected paradigm for Haf in a hypothetical dialect with reanalysis of hafia

<table>
<thead>
<tr>
<th>UNDERLYING</th>
<th>SURFACE</th>
</tr>
</thead>
<tbody>
<tr>
<td>default form</td>
<td>/hæf/</td>
</tr>
<tr>
<td>3rd person singular</td>
<td>/hæfs/</td>
</tr>
<tr>
<td>gerund participle</td>
<td>/hæf + ɪŋ/</td>
</tr>
<tr>
<td>perfect participle</td>
<td>/hæf + d/</td>
</tr>
</tbody>
</table>

Joseph (1992:137) gives evidence that might indicate the existence of dialects with a reanalysis like this; instead of having to these dialects use [hæfɪŋtʊ]. Joseph says regrettably little about these dialects, providing hardly any information about them except that they use [hæfɪŋtʊ] for having to. It would be very interesting to know what the other forms of the paradigm are for them, and what the syntax of the novel verb is (for example, whether they allow adverb interpolation between Haf and to; see §7.2 below). I do not know of any relevant speakers, so I am unable to investigate further at the moment, but it seems quite plausible that they might instantiate the reanalyzed system shown in 8. In dialects like this, the reanalysis would have severed the synchronic link between HAVE and the new necessitative verb HAF. I did not accept such a reanalysis in §3 above because there I was talking about the standard dialect, which does not seem to have severed the link, but the dialects Joseph reports may be different.

6.3. OUGHTA. The modal ought is an auxiliary verb by the usual tests (fronting in polar interrogatives, taking the -n't suffix, etc.). The phonology that yields [ɔrə] from [ɔt tu] provides further evidence of special word-internal phonology in -to derivatives. The default would be to assume that sequences like ought to be smart have a phonology paralleling sequences like thought to be smart. But in fact they differ: colloquial but careful American speech permits phonetic [ɔrə] in they oughta but not in they are thought to. The to-derivative oughta has the phonology of a single word like auto, while thought to has the phonology of a word sequence. In the latter, to must obey the general constraints on the phonology of stranded infinitival to, notably that it not be completely deaccented (Zwicky 1982).

\(^{13}\) Sag and Fodor take the ‘anomalous regressive assimilation required to derive hafita and hasta’ as evidence in favor of the separate-lexical-item position, but this is not a good argument. Regressive voicing assimilation to the inflectional ending /-z/ is found exceptionally in wife + z = wives, knife + z = knives, etc., and is not taken to indicate that the plurals are separate lexical items from the singulars in these cases.
6.4. Gotta. The same contrast can be seen in the case of necessity got.
The form orthographically represented as (gotta), morphologically got-to, phonetically [ˈɡərə], does not rhyme with the sequence not to. In standard English, necessity got only occurs as a complement to perfect have. In colloquial American English, it occurs without have as in I gotta go, with negative I don’t gotta go, polar interrogative Do you gotta go?, etc. Some speakers have zero inflection in the 3rd singular present (She gotta find another place); generally these speakers’ usage in other paradigms indicates that they have regularized many irregular 3rd singular forms to the base form (e.g. She don’t got no home to go to). Other speakers (as noted by Schmerling 1983: n. 3) inflect gotta on its head for the 3rd singular, saying She gotsta find another place, as predicted by the morphological analysis defended here (but not by the lexicalization account).

6.5. Sposta. Expectational supposed also provides evidence of special phonology in the paradigms of the therapy verbs. The form sposta or supp sta does not have the same phonology as sequences like host to: compare You should pour the champagne, you’re sposta [ˈsposta] ‘you’re supposed to’ with *You shouldn’t pour it, you should wait for the hosta [ˈhosta] ‘you should wait for the host’.

6.6. Usta. Habitual used, which combines with -to to form usta, is a non-auxiliary catenative verb serving to mark what in many languages would be a past imperfect tense and/or aspect and having only the past tense form and (for most speakers but perhaps not all) a homophonous base form seen in We didn’t usta but we do now. There is a diachronically related item meaning ‘habituated’, as in I’m not used to being the center of attention, but it has different phonology from past habitual used: the sentence I don’t know if he still does, but he used to can be pronounced in a way that ends with usta, phonetically [ˈjuːsta], even in very carefully enunciated speech; but the sentence It’s not what I’m used to ends in [ˈjuːstu], because it does not participate in the same special phonology. The constraints on stranded to (Zwicky 1982) apply to this occurrence of to, and it must have the light accent that has [tu] as its concomitant pronunciation.

6.7. Wanna. With desiderative want, the most discussed of all the therapy verbs, the claim to special phonology is less clear. The phonology involved in the reduction of want to to wanna [ˈwɑːnə] is found elsewhere, for example, in the phonology of the reduction of Santa to [ˈsæntə]. But even here, as noted by Jacobson (1982: 203, 215–16), want does not have exactly the same phonologi-

---

14 Admirers of the songs of Michael Flanders and Donald Swann may recall the lines 'You can’t pass heat from a cooler to a hotter [ˈhɒtə]/You can try it if you like but you’d far better notch [ˈnɔtə], which rhyme in British English. The joke, of course, is that the contraction of not to in you’d far better not to is illicit, and the rhyme is thus segmentally fine but prosodically desperate.

15 Lightfoot (1976) takes the transitive verb use [juːz] to be the same verb as habitual used, and builds an argument on this equation, but it is hard to see why, since these items differ phonetically, phonologically, morphologically, syntactically, and semantically.
cal behavior as words of similar phonological structure (wanted typically loses its [t] but taunted does not), and wanna shares this (we get I wanna get the font to match at slow rates of speech, but we do not get *I wanna get the fonna match).

7. THE SYNTACTIC FACTS. The syntactic facts claimed to be relevant to the correct analysis of to-contraction, on which scores of pages of argument were published during the 1970s and 1980s, can now briefly be reviewed. Remarkably, all of them fall into place quite naturally under this account, some of the facts receiving a satisfactory account for the first time. Under the morphological treatment none of the syntactic data motivates or supports any special assumption about syntactic theory. In particular, no conclusions emerge from this domain of facts (positively or negatively) about inaudibilia or 'empty categories'. Thus in the view I take here the central claim of Sag & Fodor 1994 is fully supported, as are the conclusions of the series of papers by Postal and Pullum: to-contraction phenomena provide no evidence for traces.

We can pass rapidly over the trouble that early analyses had with the fact that all the therapy verbs other than desiderative want are raising verbs (see the critique of Lightfoot 1976 in Postal & Pullum 1978); assuming NP movement, an example like There hafta be some answers would simply have the structure [there, hafta [t, be some answers]].

Difficulties relating to the AUX node (later called Infl) likewise become irrelevant; if items like oughta are morphologically derived, then issues of whether oughta is under AUX (or Infl) and what the source is for to disolve away, provided oughta is [– AUX] (the default for all verbs) and thus ineligible for inversion. Traces in intermediate complementizer positions left by successive cyclic movement (a problem first raised by Postal & Pullum 1978: 8–12) also cease to be a problem: the unwanted prediction that 9a will be impossible because of the leftmost trace in the structure 9b is not made under the morphological derivation account, where there is no to-contraction to block, and 9a has the structure 9c.

(9) a. Who do you wanna see?
   b. [who, do [you want [t, [PRO to see t]]]]
   c. [who, do [you wanna [t, [PRO see t]]]]

Some further classes of facts, however, offer somewhat more interesting confirmation of the correctness of the analysis than these first four familiar example types. These are discussed in §§7.1 to 7.8.

7.1. HEAVY NP SHIFT. Postal and Pullum (1978: 25) note that heavy NP shift does not feed to-contraction. The marginal 10a is distinctly better than 10b.

(10) a. *I want to present themselves in my office all those students who failed the test.
   b. *I wanna present themselves in my office all those students who failed the test.

Heavy NP shift is sometimes taken to be a stylistic rule that would not be expected to leave traces; this could be an embarrassment for some trace-based
accounts. The issue does not arise under the morphological account, since the derived forms like wanna are subcategorized for a subjectless VP complement (or a nonfinite cause with controlled PRO subject). Example 10b is ungrammatical for basically the same reason as 11.

(11) a. *I tried to present themselves in my office all those students who failed the test.

There is no grammatical role for the extraneous NP (all those students) in 11. Sentences like I want those students to present themselves and the sentences like 10a related to them by heavy NP shift are not relevant in any way to the syntax of sentences with wanna.

7.2. Interpolation resistance. Chomsky and Lasnik (1978) note that some people do not allow adverb interpolation at all in the sequence have to:

(12) %We have obviously to be careful.

Just because Hafta is a morphological derivative of have and takes a base-form VP complement, that does not entail that have will be subcategorized for a to-infinitive. For many people (especially older generation, and more British than American speakers) it is, and 12 is grammatical, but for speakers who reject interpolation it is not; Hafta is available but necessity have does not occur (or does not occur productively; the speakers might have some remnant expressions like If you need anything, you have only to ask, used rather in the manner of quotations from another dialect). This should be no more surprising than the absence of *[ν aggress] from the idiects of most speakers despite the presence of the noun [n[ν aggress]-ion] derived by the relatively productive morpholexical rule that yields a noun of the form [n[ν X]-ion] from a (Latinate) verb stem X.

7.3. The langendoen dialect. Postal and Pullum (1978: 27) note the curious fact that D. Terence Langendoen reports accepting sentence 13, suggesting under syntactic accounts that in his dialect to can adjoin to want even where a main clause adverb phrase linearly separates them.

(13) %I wanna very much go to the game this evening!

(The intonation pattern associated with this sentence type, as Langendoen volunteered it orally, is rather indignant-sounding, with high pitch and heavy accent on much.)

This datum poses quite a problem for many of the to-contraction analyses, since they trade on a syntactic adjacency requirement that even a movement trace will suffice to block. Here an overt adverbial phrase fails to block the alleged contraction operation. It looks as if a sequence want very much to is contracting to want + to very much despite both syntactic and phonological nonadjacency of the host and the contracting item.

Chomsky and Lasnik (1978) call the phenomenon ‘contraction at a distance’ and deem it ‘highly marked, if even possible.’ They make a valid point about the analysis advocated by Postal and Pullum (1978): it makes Langendoen’s dialect easy to describe without any constraint but needs a special linear precedence condition (that the clause-union amalgamation of want and to must
be strictly adjacent to its complement) to capture the much more widespread standard dialect. However, Chomsky and Lasnik’s own account struggles to cover the facts. They are forced to postulate ‘a dialect-specific idiosyncractic rule for Langendoen’s dialect inverting to and very much’ (p. 271). They stipulate that this rule yields the structure ‘want + to very much’, where the + is a sign of a word boundary having been reduced to a morpheme boundary, making the contraction obligatory under Chomsky and Lasnik’s assumptions. This yields the consequence that the contraction will be obligatory, which is needed because, as they report, in Langendoen’s dialect the word order in question is not available with uncontracted want to or with other infinitive-complemented verbs; thus 14 is ungrammatical for Langendoen, indicating that he does not in general allow main-clause adverbs to be placed inside embedded infinitive phrases.

(14) *I hope to very much go to the game.

Chomsky and Lasnik remark in a footnote: ‘Alternatively, it might simply be that wanna has been lexicalized, with new and specific properties’ (n. 4). The morphological account proposed here makes this idea precise: the Langendoen dialect simply allows a verb to be linearly separated from its base-form complement.

This assumption has consequences for other data in Langendoen’s dialect. For most speakers, help also takes a bare infinitival VP complement (though it is not in the same verb class as wanna: He helped wash the dishes has an implicit unspecified direct object, made explicit in He helped them wash the dishes). This suggests that 15b should be as grammatical as 15a for Langendoen.

(15) a. I help wash the dishes every night!
   b. %I help every night wash the dishes!

And indeed, Langendoen reports that (with the same intonation as 13) this is the case. Thus what is at issue is which complements of a verb may be linearly separated from it by an adverbial phrase. Infinitival VPs with to can generally be separated from their matrix verbs (essentially everyone accepts I helped every night to wash the dishes), but for many speakers, it seems that a bare VP complement of a nonauxiliary verb is as tightly constrained to be verb-adjacent as an object NP, so they will judge 13 and 15b to be ungrammatical. Langendoen differs in that for him the constraint requiring verb-adjacency for certain complements is weak to nonexistent, at least under some circumstances, when the complement is a bare infinitival VP.

7.4. Coordination. Postal and Pullum (1982) observe that when two or more verbs that take to + VP are coordinated, the last of them being want, the resultant want to sequence cannot be pronounced wanna: string 16 is ungrammatical.

(16) a. *I don’t expect, need, or wanna dance.
   b. *I don’t wanna or expect to dance.

Under many to-contraction views, this is puzzling. Notice that the rule determining choice of phonological shape for the definite article makes reference to
the consonantality of the following segment regardless of the fact that it is in a coordinate structure: we get an orange or banana with [ən] and a banana or orange with [ə]. And as Carden (1983: 42) notes, the assimilation rule yielding [tʃə] from /tʃə/ does operate across the boundaries of a coordinate structure in with or without you, which can be pronounced in a way that ends [wiðautʃə]. Chomsky and Lasnik (1978: 271) compare their rule of to-contraction with ‘external sandhi’ as if it were in a class with other low-level rules of morphosyntactically conditioned phonology, but in fact the alleged contraction does not behave in the way true sandhi rules do; Sanskrit sandhi rules do apply across the boundaries of coordinate structures.\textsuperscript{16}

Under the analysis defended here, the question of whether (or which) rules of certain types apply across the boundaries of coordinate structures becomes irrelevant. What is wrong with 16 is simply that two verbs with different subcategorizations have been conjoined. The case is analogous to that of 17.

(17) *We’ll have to persuade or make him do it.

It has nothing specifically to do with any special properties of the therapy verbs, but merely to do with facts about what are called coordinate factors in Pullum & Zwicky 1986.

7.5. Adverbial infinitival clauses. To a limited extent the verb want can be used intransitively with the meaning ‘to possess desires’ or ‘to be lacking in necessities’ (as in Waste not, want not). This permits us to test the behavior of the to of an adverbial infinitival clause positioned immediately after it. As noted in Postal & Pullum 1982, there will be no intervening trace in such a case, but just a controlled PRO subject, under modern transformationalist accounts. Thus those accounts will predict contraction. This prediction is not confirmed. Consider the data in 18.

(18) a. (In order) to be an effective over-consumer, you have to really want.
   b. You have to really want (in order) to be an effective over-consumer.
   c. You have to really wanna be an effective over-consumer.

If wanna is used instead of want to in a version of 18b without the phrase in order, the adverbial reading of the infinitive disappears completely. No one understands 18c as synonymous with 18a. Clearer evidence is available from prospective go, as shown in 19.

(19) a. I’m not going to annoy her.
   b. I’m not gonna annoy her.

Of the two readings of 19a, roughly ‘I will not annoy her’ and ‘It is not just to annoy her that I will go’, we find only the first in 19b.

None of the trace-based accounts capture facts of this sort, since under no set of assumptions is there a relevant movement trace between want and to. To deal with the problem, post-1982 accounts adopted the assumption that the

\textsuperscript{16} Thanks to Brian Joseph and Paul Kiparsky for providing textual examples that confirm this.
government relation that obtains between a matrix verb and the infinitival *to* of its complement phrase does not obtain between a matrix verb and the *to* that introduces an adverbial infinitival clause. The structures might be as in 20, for example, with the definition of government crafted to include as governees all constituents in the most embedded maximal projection of which the matrix verb is the lexical head (and to reach as far down as the T constituent where *to* is usually assumed to be located) but not to include constituents under higher maximal projections of which that verb might also be regarded as the head in some sense, e.g. the root VP in 20b.17

(20) a. Complement-clause structure

```
VP
   \---+--
      \---C
           \---IP
                   \---NP
                   \     \---I'
                      \---I
                          \---VP
```

going   to   annoy her

b. Adverbial infinitival clause structure

```
VP
   \---+--
      \---CP
           \---IP
                   \---NP
                   \     \---I'
                      \---I
                          \---VP
```

going   to   annoy her

But the real defect of any such account is that it makes it merely a stipulation that the constraint is couched in terms of ‘must govern’ rather than ‘must be governed by’ or ‘must be in a mutual government relation with’ or ‘must not govern’ or any other relation definable in phrase markers. Under my analysis here, the existence of sentence types with adverbial infinitival clauses modifying the matrix VP, or sentences of any sort with want, cannot affect sentences containing wanna.

7.6. Other accidental adjacencies. Postal and Pullum (1978, 1982) present a range of other examples designed to force an instance of want to occur

17 To be precise, government would have to be defined using the strict succession version of MAX-command; see Kornai and Pullum (1990:44–6). And under the usual definitions of government, which do not imply that if A and B then A governs all the nodes dominated by B, it would also have to be assumed that projections not dominating any terminal materials exclusive of other projections—such as the CP and C nodes in 27—are pruned or in some other way made not to count.
accidentally adjacent to an instance of to that does not introduce its infinitival complement: by using want as a noun before an infinitive as in 21a, using (albeit very awkwardly) a subjectless infinitival clause as a sentential subject of the complement of want as in 21b, or in various other ways.

(21) a. We cannot expect that want to be satisfied.
   b. *I don’t want [(to do nothing] to be the response of this administra-
Postal and Pullum point out that the pronunciation wanna is never permitted for such want to sequences. In each of these cases, the same remarks apply as have been made in previous sections: the data are fully explained under the assumption that there is no rule of to-contraction to apply or be blocked from applying. The morphologically derived item wanna has the verb want as its head, but has a distinct subcategorization from want (just as overcome has a distinct subcategorization from come): it needs an immediately following bare infinitival subjectless complement. The various contexts in which the form to can be made to fall adjacent to an occurrence of want immediately to its left are not relevant to the distribution of the to-derivative verb wanna.

7.7. Phonological phrasing. Postal and Pullum (1978) point out that if all to-contraction amounts to is a syntactic rule saying ‘want to → wanna’ that is blocked only by intervening syntactic material such as traces, there is no explanation of why it cannot apply across the boundaries of an interpolated parenthetical phrase, yet this produces extreme ungrammaticality.

(22) a. I want, to be precise, a yellow four-door De Ville convertible.
   b. *I wanna, be precise, a yellow four-door De Ville convertible.

In response to this objection, Chomsky and Lasnik (1978: 271) suggest that ‘surely any reasonable theory of contraction will be restricted to within an intonational phrase, a fact that we did not trouble to mention.’ But far from being an obvious codicil not needing comment, this idea is entirely unexplicated by Chomsky and Lasnik, and in fact appears to be incompatible with any syntactic account of contraction. Syntactic rules are not sensitive to phonologi-
cal or phonetic distinctions. Rather, phonological phrasing is assigned to syntac-
tic structures that are independently defined as well formed. There is nothing in general linguistic theory that explains why the contraction rule could not apply and the intonational phrasing be assigned as shown by the commas in 22b. Syntactic rules cannot be allowed to be sensitive to intonational phrase boundaries without predicting a host of never-attested phenomena (see Zwicky 1969, Zwicky and Pullum 1986, Miller, Pullum and Zwicky 1997, and many other works on the principle of phonology-free syntax). It is an essential fact about syntactic processes that they determine the distribution of elements of sentence structure without reference to morphological or phonological factors.

Under the account I propose, however, not only is this issue dissolved, but phonological phrasing turns out to be irrelevant to the ill-formedness of 22b.

\footnote{The term intonational phrase was not in use in 1978 phonology; it seems to originate in work by Lisa Selkirk starting from around that time (see Selkirk 1984 for references).}
It is certainly the case that intonational phrase boundaries cannot occur in the middle of words, but that is not what is wrong with 22b. If *wanna* is a morphological unit, the infinitival phrase *be precise* in 22b lacks the infinitival *to* that is an obligatory part of the construction, and in addition, *wanna*, which is subcategorized for a bare infinitival VP complement, occurs with a direct object NP. The pair in 22 is thus parallel to the pair in 23, and it is no surprise that 22b is as starkly ungrammatical as 23b.

(23) a. They wooed, to be precise, every member of the opposite sex they encountered.

   b. **They would, be precise, every member of the opposite sex they encountered.**

In 23a the verb *wooed* has an NP object as is appropriate, and the parenthetical infinitive phrase has *to*; while in 23b, the verb *would* fails to have the bare VP complement it is subcategorized for and the parenthetical phrase lacks the *to* that is syntactically demanded. There is no need here for any syntactic constraint calling for reference to intonational phrase boundaries.

7.8. THE LIBERAL DIALECTS. Finally, I consider the implications of the liberal dialects, which do not exhibit the *Teddy* facts. The later trace-theory accounts of *wanna* attempted to deal with the problem of NP-movement traces, traces in complementizer positions, and other problems by motivating the idea that Case-marked traces universally block contraction, whereas PRO and non-Case-marked traces are universally transparent to phonological processes. But it is by no means a fact of universal grammar that *wanna* is not found as the pronunciation of *want to* sequences that would be separated by a Case-marked trace under modern transformationalist assumptions. There exist dialects in which contraction is systematically permitted across the alleged Case-marked traces, so that strings like 24, ungrammatical for me and many others, are grammatical, with the meaning ‘Who do you want the driver of the car to be?’

(24) *%*Who do you wanna drive the car?

It is here, rather than in examples like 22 where Chomsky and Lasnik invoked it, that phonological phrasing does have a role to play. I claim that liberal dialect speakers are those who permit the phonological rules reducing *want to* to *wanna* to apply phrasally. I suspect (though this is a matter that would repay further research) that it applies within an intonational phrase in the sense of Selkirk 1984.

Notice that the *want to* sequences in Postal & Pullum 1982 that most sharply resist substitution by *wanna* are precisely those where it is thoroughly unnatural to phrase the words together phonologically. The facts of 22 illustrate nothing for dialects that do exhibit the *Teddy* facts, as we have just seen, but for liberal dialects they clearly indicate that indeed contraction is an intraphrasal phenomenon: even speakers who judge 24 to be grammatical do not come anywhere near finding 22b tolerable.

But consider also examples like 21b, repeated here as 25, where *want* and *to* are adjacent but the *to* is in a (syntactically awkward) sentential subject of a subordinate clause:
(25) I don’t want [(to do nothing) to be the response of this administration]. I know relatively little about the liberal dialects (my contacts are nearly all American speakers for whom the Teddy facts are real, or British speakers who do not really find wanna a natural item in their speech), but I know of no speakers who would allow replacement of want to by wanna is a case like 25, and I doubt that there are any. But notice, it is also quite difficult to envisage 25 being pronounced in a way that phrases to do nothing with I don’t want. The sentence is cumbersome at best, but if it were to be used at all the speaker would surely phrase it as shown by the curly brackets in 26.

(26) {I don’t want} {to do nothing} {to be the response of this administration}.

In a case where wanna is admissible, it is always possible, even preferred, for want and to be intonational phrase mates, as in any of the phrasings in 27.

(27) a. {Perhaps he doesn’t want to do it}
   b. {Perhaps he doesn’t} {want to do it}
   c. {Perhaps} {he doesn’t want to do it}

On a similar basis, I predict that liberal dialect speakers will be most unlikely to favor using wanna in a case like the purpose-clause reading of 28.

(28) You have to really want to be an effective over-consumer. That is because the only natural intonational phrasing, the one that brings out the separation between the main clause and the adverbial clause, would be the one in 29.

(29) {You have to really want} {to be an effective over-consumer}

For the liberal dialect speakers, the phonological shape spelled ⟨wanna⟩ is phonologically derivable from the sequence want to in contexts where those words occur phonologically adjacent within the boundaries of a single intonational phrase.19

This is the point on which the Postal and Pullum (1978) syntactic analysis founders. Postal and Pullum say nothing to suggest a reason why liberal dialect speakers, if they simply lack the clause union condition on pronouncing want to as wanna, would not accept examples like 22b, or the adverbial reading of 18c, or the second reading of 19b, or the examples of 21 with wanna, which I confidently predict they do not. The analysis I offer predicts there will be three types of dialects: those with a morpholexical rule producing to-derivatives but no phonological reduction of want to to wanna; those with a phonological reduction of want to to wanna that operates within the boundaries of intonational phrases only; and (possibly) those with both.

8. Conclusions. In 1970, when the Teddy facts were first discussed in this journal, they were claimed to show that nonsuperficial syntax—facts about earlier stages in transformational derivations—could influence phonology, and thus that linguistic theory had to countenance global derivational constraints.

19 Whether or not such speakers also have the morphological derivation rule forming wanna would not be easily determined. If they do, they maintain a dual analysis for sentences containing wanna.
Allegations about excessive descriptive power and arbitrariness of constraints were subsequently hurled in various directions. One theoretical principle at stake was what Zwicky (1970) called the PRINCIPLE OF SUPERFICIAL CONSTRAINTS IN PHONOLOGY (PSCP), which says that where phonological rules make reference to properties of syntactic structure, they refer only to surface syntax. The PSCP is respected a fortiori in the account advocated here: far from there being global syntactic constraints linking nonsuperficial syntax with phonology, there are no special syntactic constraints at all.

Post-1970 literature has muddied the picture of the PSCP by providing surface syntactic indications of nonsuperficial constituent positions, i.e. traces. The apparent syntactic constraints on to-contraction were widely taken to support trace theory. All such support is removed under the present analysis.20 Although the syntax and phonology of the therapy verbs has served as the basis for dozens of contributions to a rhetorical battle over syntactic theory, and has masqueraded as a prime argument for the reality of movement traces, if this my analysis is correct the phenomenon is simply an illustration of how morphological facts can be mistaken for syntactic ones.21

The error in the syntactic treatments that posited the supposed rule of to-contraction is at root one of setting defaults the wrong way round. There has been an unquestioned assumption that a general rule attaches to to an item on its left; the default was that the rule should be permitted to apply. When an appropriate item (such as WANT) occurred immediately before to but contraction was not permitted, an explanation was needed. This set linguists looking for general constraints on contraction: that it could not occur across a trace, or across a Case-marked trace, or between an instance of to and an occurrence of WANT that did not govern it. But the search was a wild goose chase. There is no general syntactic process of contraction or adjunction to apply or to be blocked. Forms like wanna (in the standard dialects) occur only as occurrences of to-derivatives produced by lexeme-to-lexeme morphological rules that assign them specific subcategorization properties. No general conclusions about constraints in syntax or in phonology are supported by the syntactic distribution of wanna or the other to-derivatives of the therapy verbs.

APPENDIX

It ought to be possible to bring new evidence to bear on the correctness of the analysis defended here by considering facts about the ellipsis and fronting of VP complements, but the matter is complex and turns out to yield nothing conclusive. In this appendix I briefly outline the issues.

20 It has also been widely believed, since King (1970) introduced the idea, that the distributional conditions on the weak forms of function words in English (e.g. the condition that allows I wonder where it is now? but blocks *I wonder where it's now?) provide support for traces. Pullum and Zwicky (1997) show that this too is false.

21 It is not the first such example observed in English. As foreshadowed in Zwicky (1970, 327–28, n. 7) and Palmer (1974: 237ff), argued in detail by Zwicky and Pullum (1983) in this journal, proposed independently Huddleston (1984: 87f) around the same time, and independently arrived at by Williams (1994), the n't of negated auxiliary verbs is actually an inflectional affix, not an enclitic, and not (synchronously) a reduced or contracted form of the negative particle not.
It is well known that VP ellipsis demands a [+AUX] element as the left context. (The way I would like to put it, a [+AUX] VP may be nothing more than a lightly accented lexical head, the semantics being obtained by applying the lexical head to a free variable over VP meanings that is bound to a linguistic antecedent in the discourse.) VP ellipsis is thus allowed in (a) and (b), but not in (c) or (d) where verb phrases have been omitted after nonauxiliary catenatives.

(a) I already [VP[-AUX]] have.
(b) He probably [VP[-AUX]] will.
(c) *They really [VP[-AUX]] seem.
(d) *Most people [VP[-AUX]] tend.

Infinitival to behaves like a nonfinite [+AUX] verb; its lack of tense ensures that it will not be found in subject-auxiliary inversion contexts or affixed with n't, but it does provide VP ellipsis contexts:

(e) Most people [VP[-AUX]] tend [VP[+AUX]] to[].

But wanna is not [+AUX]; although it can occur tensed, it does not invert in polar interrogatives. Thus my analysis does not immediately predict that we will find data like these:

(f) I didn’t go because I didn’t wanna.
(g) I hafta and I ain’t gonna. (Robert Redford’s line from the film Butch Cassidy and the Sundance Kid)

One might think I could simply conclude that the to-derivatives of the therapy verbs have been added to the set of left-context elements for VP ellipsis, but this would not be correct. The feature defining the left context for VP ellipsis is [+AUX], and as the usual tests show (note *Oughta we go?, etc.), none of the therapy verb to-derivatives have that feature. And the therapy verbs do not constitute a syntactic class sharing a feature like the [+AUX] verbs, but merely a class of lexemes that are the domain for a specific morphological rule. Assuming (as I do) a theory in which syntactic conditions cannot make reference to purely morphological or phonological information, making reference to the therapy verbs in a syntactic condition is no more allowable than making reference to those nouns that are formed by the suffix -ment.

However, there is another regularity in English grammar permitting complement ellipsis: the one Hankamer and Sag (1976) refer to as NULL COMPLEMENT ANAPHORA (hereafter NCA). Whether or not a verb permits NCA is clearly a lexical matter. As illustrated in (h) VOLUNTEER and TRY permit it, while the verbs ARRANGE and ATTEMPT do not.

(h) They couldn’t persuade anyone to collect the tickets, so Susan volunteered.
(i) I may not ever have read Finnegan’s Wake, but I’ve tried.
(j) *They couldn’t persuade anyone to collect the tickets, so Susan arranged.
(k) *I may not ever have read Finnegan’s Wake, but I’ve attempted.

Assuming that the to-derivatives of the therapy verbs are NCA verbs would account for the existence of examples (f) and (g).

It would be quite natural for the to-derivatives of the therapy verbs to become NCA verbs, of course. The children in the first generation to introduce the to-derivatives into the language, having heard a wanna-like pronunciation for want to and posited a lexical relation between wanna and want, would still hear adults saying things like I don’t want to with a wanna-like pronunciation. As the evidence established the relevant generalizations about VP ellipsis and NCA, wanna would naturally be taken to fall under the latter generalization if it did not fit the former.

The trouble with using the status of ellipsis as a new test of my analysis is that it is quite hard to find reliable evidence that NCA rather than VP ellipsis is operative in a particular case. Hankamer and Sag note a number of contrasts of behavior between VP ellipsis and NCA, but all of them depend on data associated with significant interspeaker disagreement. One claimed difference is that VP ellipsis (a species of what Hankamer and Sag call surface anaphora) cannot be exophoric—that is, it must have a linguistic antecedent—while this is not true of NCA. They claim that I’ll volunteer, like I’ll do it, can be used with no linguistic antecedent for the missing complement, but I will cannot. See Schachter 1977 and Hankamer 1978 for a dispute about the facts.

A second claimed difference is that NCA can have a linguistically discontinuous antecedent while VP ellipsis cannot: The tickets needed to be collected, and Susan volunteered is grammatical.

---

22 I am grateful to Eric Poitdam for pointing this out to me.
but replacing volunteered by agreed to renders the example ungrammatical. But as Philip Miller has pointed out to me, it is in fact not at all unusual to encounter VP ellipsis with discontinuous antecedents in everyday language use, both written and oral.

A third difference is that missing antecedents can be found in VP ellipsis cases but not NCA cases. Hankamer and Sag claim that They said that some of us would have to give up their seats, and I did so because it was too narrow for me anyway is grammatical, but with volunteered instead of did so it is not. As Hankamer and Sag acknowledge (1976:413), ‘Missing antecedent judgments are admittedly delicate.’

A fourth contrast involves the disappearance of ambiguities in belief contexts: Hankamer and Sag state that in Susan claimed that the earth is larger than it is, and so did Harry we get a sensible reading as well as one that impounds a contradictory belief to Susan and Harry, but if so did Harry is replaced by Harry agreed, the sensible reading disappears.

For all of these cases, it is possible to construct examples that contrast want to (which involves VP ellipsis) and wanna (which under my account must lose its complement by NCA). But after carrying out such constructions, I have found that there seem to be no really reliable intuitions about their status. Speakers are uncertain enough about the contrasts above anyway, and the contrast between want to and wanna makes it even harder to get judgments. A few speakers have told me they find wanna worse than want to in crucial cases, which would support my analysis, but much as I would like to believe it, I have decided I simply cannot be sure of any of the contrasts, so I will not take the space to present the examples here.

Paul Postal suggested to me that VP fronting might offer some evidence. This rather rare and unwieldy construction is possible only for VP complements to auxiliary verbs, as seen by the contrast of (l) and (m).

(l) I said I’d wash the dishes, and wash them I did.

(m) *I said I’d help wash the dishes, and wash them I helped.

We would therefore expect to-derivatives of therapy verbs not to allow VP fronting of their complements. The trouble is that to construct the relevant contrasts we need to compare VP fronting after wanna with VP fronting after to, and this is not possible: VP fronting appears to be limited to VPs following finite auxiliary verbs. Not a single one of the examples of VP fronting in Ward 1990, the most data-intensive study of the construction I have seen, involves fronting of a VP from a nonfinite context such as after infinitival to. My attempts to question informants about whether they perceived a distinction between pairs like climb it I want to and climb it I wanna produced extremely ragged and inconsistent results, and revealed nothing that could be judged either a positive or a negative result for my analysis.

Sandra Chung has pointed out to me the potential relevance of a further class of facts, involving antecedent-contained VP ellipsis, as in (n), and here the facts seem slightly clearer.

(n) I’ll eat anything that you will.

Ex. (n) is paraphrasable as ‘I’ll eat any x such that you will eat x’. NCA does not permit such containment of the missing complement inside its antecedent. Demonstrating this is not that easy (many NCA verbs take NP objects as well as clausal or VP complements, and this can lead to other irrelevant readings for the crucial example), but it seems to be possible; (o) is quite clearly not an acceptable way to express the proposition ‘I always eat any x such that he volunteers to eat x’.

(o) *I always eat anything that he volunteers.

What makes this relevant for me is that with VP ellipsis after want to we get acceptable examples like (p) but the result of substituting wanna to get (q) is much worse.

(p) I always want to eat anything that you want to.

(q) *I always want to eat anything that you wanna.

If this contrast is real, it supports my analysis by providing evidence that omitted VPs after to-derivatives are missing by virtue of NCA, not VP ellipsis. The result, however, does not strike me as very secure, and I feel the topic needs further research.

REFERENCES

Schmerling, Susan F. 1983. Another perspective on wanna. Austin, TX: University of Texas, MS.

Stevenson College
University of California, Santa Cruz
Santa Cruz, CA 95064
[pullum@ling.ucsc.edu]  [Received 6 February 1996; revision received 3 September 1996; accepted 5 September 1996.]