Subject Ellipsis in English: Construction of and Findings from a Diary Corpus

Laura Teddiman[1], John Newman[2]
University of Alberta

Abstract

Subject ellipsis is not typically considered to be a common occurrence in English, although it has been studied within certain genres of text, including personal diaries. The current paper explores English subject ellipsis in a diary corpus that has been constructed from online weblogs (blogs). The pronouns *I* and *it* were found to be the most frequently omitted subjects and their identities were usually recoverable from preceding linguistic information. Both main verbs and auxiliary verbs were used by authors in sentence initial position, with HAVE, BE, and GET being the most common. Also discussed is the privacy of these personal, yet public, journals. Results are similar to those found for studies of subject ellipsis in casual letters, but are suggestive of differences based on authorial intentions and the intended audience.

Keywords : subject ellipsis, diary texts, weblogs, corpus.

1. Introduction

Sentences in English typically contain overt subjects. However, subject ellipsis is not a rarity, especially in conversation (Biber *et al.* 1999: 1048). They are often easily interpretable and do not appear to be errors on the part of speakers. For example, "don't know" is an understandable reply when responding to a question one does not know the answer to, even though the subject "I" is omitted. Subject ellipsis is also attested in certain textual registers such as diary registers (e.g., Haegeman and Ihsane 1999) and in telegraphic communiqués where economy is paramount (Barton 1998). The current paper will briefly discuss previous corpus based research of English subject ellipsis before turning to a discussion of a diary corpus constructed from online journals.

While there have been theoretical attempts made to explain English subject ellipsis, they have not generally described the phenomenon in great detail. Most have addressed the issue of ellipted subjects from within a Generative framework, with particular reference to the *pro*-drop parameter. Within studies of child language acquisition, there has been an emphasis on the frequent omission of subjects in child speech, even in languages that are considered to be non-*pro*-drop languages such as English (Valian 1990; Valian 1991). Such studies tend to focus on the complexity of the phenomenon and the development of syntactic parsers in children.

---

[1] Department of Linguistics, University of Alberta, teddiman@ualberta.ca

[2] Department of Linguistics, University of Alberta, john.newman@ualberta.ca

Within adult populations, the presence of null subjects in diary texts has been addressed. Haegeman and Ihsane (2001: 329-331) suggest that *pro*-drop and null subjects found in diary registers are comparable to the *pro*-drop stage in language acquisition. Results from this line of research suggest that most diaries are written in a specific diary dialect, where subject omission is allowable for main clauses (e.g., *went to the store*). This work builds on Haegeman and Ihsane (1999), which considered subject ellipsis inside embedded clauses in diary registers, and discussed the difficulties posed to the empty category theory. They propose that, because English speakers cannot use verbal agreement to identify ellipted subjects, they must look to antecedents in the broader context of the text, but do not make any claims about the syntactic properties of a null subject that might influence its identification. However, some of the diaries used in the construction of their corpus were taken from commercially published material (e.g., the book *Bridget Jones's Diary*) and may not have been entirely representative of ordinary writing.

More recently and with greater emphasis on the contextual surroundings of null subjects, Nariyama (2004) examined a corpus of Australian English composed of dialogue from Australian television dramas, three recorded conversations, and three casual letters, and argued that subjectless sentences in English carry different connotations than their 'normal' equivalents. Within the casual letters, *I* was omitted in 14 of the 17 instances where subjectless sentences were observed. *It* was the omitted pronoun in the other three cases. Although they were present in spoken dialogue, there were no examples of *you, s/he,* or *they* being omitted in the written text. It is suggested that because the text is written, the first person *I* is more likely to be assumed by a reader. It is also suggested that the omission of subjects in emails reflects a more casual attitude towards the reader on the part of the author. In all cases, the omitted subject could be recovered from the surrounding context. Key to Nariyama's (2004) study is the claim that subjects should only undergo elision when they can be recovered by using information in the surrounding text. Four triggers of subject ellipsis are identified: anaphoric deletion, where the subject is typically present in the preceding sentence and can be reconstructed based on linguistic information; deixis, where the subject must be reconstructed based on the surrounding non-linguistic context; dummy *it* deletion, where a non-referential "it" is deleted; and the use of conventional expressions (e.g., *gotta go*, *dunno*). The purpose of the current paper is to further explore subject ellipsis in English blogs, in particular in personal journals.

2. The Diary Corpus

The investigation began with the construction of a diary corpus. Diary text, while written, can be considered to be a representation of the internal state of the author, forming a bridge between more standard forms of writing and spoken dialogue. It may more closely approximate speech than other forms of writing and can be easily examined with available concordance software. Online diaries were chosen because the material was recent, with writers active at the time of data collection, and because the registers were informal when compared to other material posted online (e.g., news reports). Personal blogs were preferred over professional or political blogs for the same reason. Online journal entries differ from private offline journal entries in that some audience for otherwise personal entries can be assumed, if small. Writers of online diaries do not have the same assumption of privacy that an author writing to one's self does. Authors of personal blogs may presuppose an audience because their texts are posted in freely available online forums, but unlike professional bloggers, they do not tend to write persuasively towards a targeted audience for a specific purpose.

2.1. Construction of the Diary Corpus and Search Methods

The initial diary corpus was constructed using data (diary entries) available from a website hosting online blogs (Livejournal, http://www.livejournal.com). Users were randomly selected by using a regional search for recently updated diaries in the US and UK. The most recently updated diaries were used for each region. There are 50 users (writers) in each of the US and UK subcorpora, with approximately 2000 words per user (Overall total: 204,997; US: 102,781 total words; UK: 102,216 total words). Subject ellipsis was only considered in the main clause and not in coordinating structures, and whole phrase omission was not considered (e.g., *went to the store* was included in the analysis, while *to the store* was not). Age varied from 16-63 across both populations (Mean age: 25).

Data were gathered from the available websites into text files (.txt), which were used as the basis for corpus searches. AntConc (Anthony, 2006) was used to search for verb initial finite sentences. Since this was a diary text, and therefore written, initial searches were completed by inputting a period or an exclamation mark followed by a space (". " or "! ") as a search term. This revealed the beginnings of sentences that followed previous sentences, and results were then sorted by the following item (i.e., the word following the period marking the end of the last sentence). Sentences beginning in verbs were selected for further study. Further examination of the text was completed without the aid of automatic processes.

3. Results and Discussion

There were attested examples of subject ellipsis in both the US and UK dialects of English[2]. While this was not unexpected given previous work, it was important that sufficient data could be collected for analysis. In total, there were 235 interpretable cases of subject ellipsis found in the corpus in main finite clauses. The following section will present and discuss results from this corpus, with reference to Nariyama's (2004) triggers for subject ellipsis.

3.1. Reconstructed Pronouns

First person singular *I* was omitted with the greatest frequency, with 163 clear cases of subject ellipsis (68% of all cases). *It* was the next most frequently ellipted pronoun, with 38 instances (16%). The pronouns *we, she, he, they,* were ellipted in the text, but to a far lesser extent (see Table 1).

Pronoun	I	We	You	He	She	It	They	TOTAL
Observed total	163	8	16	4	4	38	2	235

Table 1: Identities of omitted subjects by frequency

In the majority of the cases, the ellipted subject could be reconstructed by examining the linguistic content of surrounding sentences, that is, the referents had undergone anaphoric deletion. Of these, 139 immediately followed sentences containing overt subjects that provided linguistic context for reconstructing the pronoun (e.g., Well, I am not in Thailand.

[2] British writers produced a slightly higher number of subjectless sentences than American writers and also used a relatively higher number of the auxiliary verbs HAVE and BE. This is especially true of the first person singular *am* (e.g., (I) Am still feeling unusually happy), which occurred five times in the British sub-corpus and was not present in the American sub-corpus. However, the differences between the two dialects were otherwise marginal, if suggestive. Potential dialectal differences could be explored using additional corpora in different genres.

(I) decided yesterday to delay the trip…). Thirteen sentences showed quasi-right deletion (e.g., (I) hope the pictures say more than **I** did), where the ellipted subject was used later in the same sentence. Another 33 pronouns could be reconstructed based on the linguistic context within two to four sentences of the same diary entry. The first person singular *I* accounts for 131 instances of anaphoric deletion and the first person plural *we* for just five. In the third person, all instances of *he*, *she*, and *they* are accounted for by anaphoric deletion, with overt referents available from linguistic context. In the second person, *you* is ellipted 7 times through anaphoric deletion. These results are similar to those described by Nariyama (2004), although the occurrence of *you* deletion is higher. However, five of the instances in this corpus come from the same paragraph (…all of you have moved on. (You) found other people. (You) found other lives …). *It* presented a more challenging case because the deleted *it* could refer to a noun or pronoun (1a) (10 instances), or *it* could refer to an entire preceding sentence (1b) (15 instances) or the topic of a sentence (1c) (7 instances).

(1) a. My hair is black with a purple sheen. (It) looks really cool.

b. I actually cleaned quite a bit of my living room up, filling up a whole gigantic trash bag with stuff for which I no longer have a use. (It) Made me feel good and horrified at the same time

c. This time next week.... I'll be living in East London!! Most of the weekend was spent packing - all books/cds/dvds/ornaments are packed. Still to be packed are all my kitchen stuff and clothes. (It) Should be ok...

I and *it* were the only ellipted subjects reconstructed in sentences containing conventionalized expressions, with 17 and 4 instances, respectively. Conventionalized expressions included items such as *thank you*, which occurs 8 times (e.g., (I) thank you dearly), *gotta* (e.g., (I) still gotta sort out a person to live with), and *turns out* (e.g., (It) turns out I have not … missed the deadline). There were only two instances of *it* being deleted when acting as an empty subject (dummy *it*) (e.g., Guess it's just that time of year, it always comes around this time. (It) Doesn't help that I went and saw Crank today.)

Situational context, and not linguistic context, is what informs elision of the deictic type. Twelve total subjects relied on situational context to be reconstructed, all of them in the first person. These examples typically relied on knowledge that a diary is written by one person (*I*) (e.g., (I) Ran out of time last night, will color/post later).

Eleven subjects could not be reliably reconstructed. In Example 2, the paragraph entry does not contain any overt pronominal subjects. There is a reasonable expectation that the author is involved in the actions described because it is a personal blog. However, because no distinction is made between singular or plural in the form of overt pronouns, and because proper names are used and then not given anaphoric referents, it is difficult for an outside reader to determine the correct ellipted subject in some cases. For example, the ellipted subject in *Pissed off bartender* could be *I* (as the writer of the blog), *we* (the author, Bec, Imogen, and/or the band), or *they* (the band, referenced in the previous sentence). Multiple referents make it difficult to use linguistic context to determine the missing subjects, and situational context does not provide additional information, as there is no further mention of the event.

(2) Thurs: Bec still here ... Imogen came over … (We) Went to see the Mercers (Ben, Kalim, etc), play one fo [sic] their last gigs. (We/I/They) Pissed off bartender. Ali brought Thomas Dennis along,

> (they/we) Sat. (We/they) Drank. (We/they) Talked. (I/we) Spent ages with bec looking for busstop [sic] because sas is really dopey sometimes.

In cases like this, the subject that has undergone ellipsis does not appear to be easily recoverable, counter to Nariyama's (2004) findings within casual letters and dialogue. This difference may be related to the type of writing in the diary style. If the target audience is oneself or a few friends, then the ellipted text may be recoverable based on personal knowledge that is unavailable to the outside observer.

3.2. Verbs

Writers did not appear to have a preference for either auxiliary or main verbs in combination with subject ellipsis. Seventy-seven auxiliary and modal verbs began sentences, with the most frequent being HAVE (22), BE (17) and DO (16). There was a preference among DO items to be negated (*didn't* 5, *doesn't* 3, *don't* 8; *did* 2). This may be because do is carrying the negation for a following main verb (e.g., (I) don't want one). *Did* also contains one of the two observed dummy *it* elisions ((It) doesn't help that I went and saw Crank today) and the conventionalized expression *don't know* ((I) don't know what I will do just yet). In contrast, BE and HAVE show relatively fewer instances of negation, with *was* (6), *had* (11) and *have* (6) being the most common. Taken together, there is a slight preference for the past tense, which may be related to the style of writing, where authors tend to describe something that has already occurred. The most frequently occurring modal verb was CAN (9), which showed a preference for negation in the present tense (e.g., (I) can't find it anywhere), and was usually used to indicate something that the author was unable to do at the time of writing.

The most common main verbs used to begin subjectless sentences were GET (25) (e.g., (I) Got a red tartan headscarf), GO (8) (e.g., (I) Went home), THANK (14), LOOK (5), LOVE (5) (e.g., (I) love that mall), MAKE (6) (e.g., (I) made plans to go to breakfast with her), and SEE (8) (e.g., (I) saw my friend Misty tonight). Interestingly, the conventional expression *looks like (it)* does not occur in this corpus. Rather, all instances of LOOK follow noun phrases in previous sentences, with the ellipted subject corresponding to the noun phrase (e.g., (it) looks a bit hard, where *it* refers to a video game). Meanwhile, the verb THANK occurred at the beginning of a sentence 14 times in the conventionalized expressions *thank you* or *thank god/goodness*. In the whole corpus in all position, *thank* occurred 37 times and in only 5 instances was it not used in one of these conventionalized expressions.

3.3. Text Formality

Nariyama (2004: 258) states that an utterance needs to be in a casual register and that the topic of conversation must be casual for subject ellipsis to be supported. The texts collected in the current research came from publicly available personal weblogs wherein the authors appeared to assume some familiarity with their readers, and were generally quite informally written. Although there did not appear to be a pattern for a reduced incidence of subject ellipsis in the discussion of less casual topics (e.g., work, illness), this may be a function of the online diary genre. Given that these texts are made available online, there is no true expectation of privacy. This is unlike casual letters or personal offline diaries, for which there is an intended audience and no expectation of outside observation. With no expectation of privacy, authors may not be inclined to discuss what they consider to be more formal or more personally charged topics. It is also possible that in (2) and cases like it, a degree of privacy can be retained by the omission of overt subjects. One who does not already have some

knowledge of the author is less likely to be able to fully interpret such journal entries. In this way, authors may exhibit some control over readership. Different patterns might be observable in weblogs written for different audiences or for the discussion of specific topics (e.g., law). Similarly, semi-professional and professional bloggers might behave differently from the average diary writer.

4. Conclusions

This study has examined the phenomenon of subject ellipsis in English by studying its occurrence in a small online diary corpus. Results indicated preferences in pronominal deletion and preferences in identities of initial verbs. The subject *I* was omitted with the greatest frequency. *I* can usually be reconstructed with minimal difficulty by readers given the context (linguistic & situational). *It* was also omitted relatively frequently. Other third person pronouns (*he, she*, *they*) were rarely omitted, as has been reported elsewhere (e.g., Nariyama 2004). Some verbs appear to be more likely to be involved in subject ellipsis than others, particularly auxiliary verbs such as HAVE and verbs used in conventionalized expressions, such as *thank you*. Text type appears to influence verb choice, in particular showing a preference for the past tense when authors record events. Results support previous investigations citing surrounding context and recoverability as important factors in decoding the identity of omitted subjects. The four triggers defined by Nariyama (2004: 250-252) were largely successful in categorizing the data. Patterns of subject ellipsis generally corresponded to those observed in other genres. However, difficulty in determining ellipted referents in some contexts suggests that personal diaries, even when made publicly available, are not written in the same style as casual letters. This study did not differentiate between texts based on the number of readers of each journal (e.g., through site statistics), but future studies may be able to explore differences in writing associated with the size and nature of audiences of personal blogs.

References

ANTHONY L. (2006), «AntConc 3.1.2 (Windows) [Computer Software]», Retrieved from:

http://www.antlab.sci.waseda.ac.jp/software.html

BARTON E. (1998), «The grammar of telegraphic structures: sentential and nonsentential derivation», in *Journal of English Linguistics* 26/1: 37-67.

BIBER D., JOHANSSON S., LEECH G., CONRAD S., and FINEGAN E. (1999), *Longman Grammar of Spoken and Written English*, Pearson Education Limited, Essex, UK.

HAEGEMAN L. and IHSANE T. (1999), «Subject ellipsis in embedded clauses in English», in *English Language and Linguistics* 3/1: 117-45.

HAEGEMAN L. and IHSANE T. (2001),«Adult null subjects in the non-pro-drop languages: Two diary dialects», *Language Acquisition* 9/4: 329-346.

NARIYAMA S. (2004), «Subject ellipsis in English», in *Journal of Pragmatics* 36: 237-264.

VALIAN V. (1990), «Null subjects: A problem for parameter-setting models of language Acquisition», in *Cognition* 35: 105-122.

VALIAN V. (1991), «Syntactic subjects in early speech of American and Italian children», in *Cognition* 40: 21-81.