Torres Cacoullos, Rena and Catherine E. Travis. To appear. *Linguistics*.
(August 2017)

Variationist typology: shared probabilistic constraints across (non-)null subject languages*

Abstract

A key parameter in received classifications of language types is the expression of pronominal subjects. Here we compare variation patterns in conversational data of English—considered a non-null-subject language—and Spanish—a well-studied null-subject language. English has a patently lower rate of expression (approximately 3% unexpressed 1sg and 3sg human subjects vs. 60% in Spanish). Despite the stark difference in rate of expression, the same probabilistic constraints are at work in the two languages. Contrary to popular belief, VP coordination is neither a discrete nor a distinguishing category of English. Instead, a shared constraint is linking with the preceding subject, a refinement of accessibility to include, alongside coreferentiality, measures of structural connectedness—both prosodic and syntactic. Other shared constraints on unexpressed subjects are coreferential subject priming (a tendency to repeat the form of the previous mention) and lexical aspect (reflecting the contribution of a temporal relationship to subject expression). Where the languages most differ is in the envelope of variation. In English, besides coreferential-subject verbs conjoined with a coordinating conjunction, unexpressed subjects are limited to prosodic initial-position in declarative main clauses, a restriction that is absent in Spanish. We propose that the locus of cross-language comparisons is the variable structure of each language, defined by the set of probabilistic constraints but also the delimitation of the variable context within which these are operative.

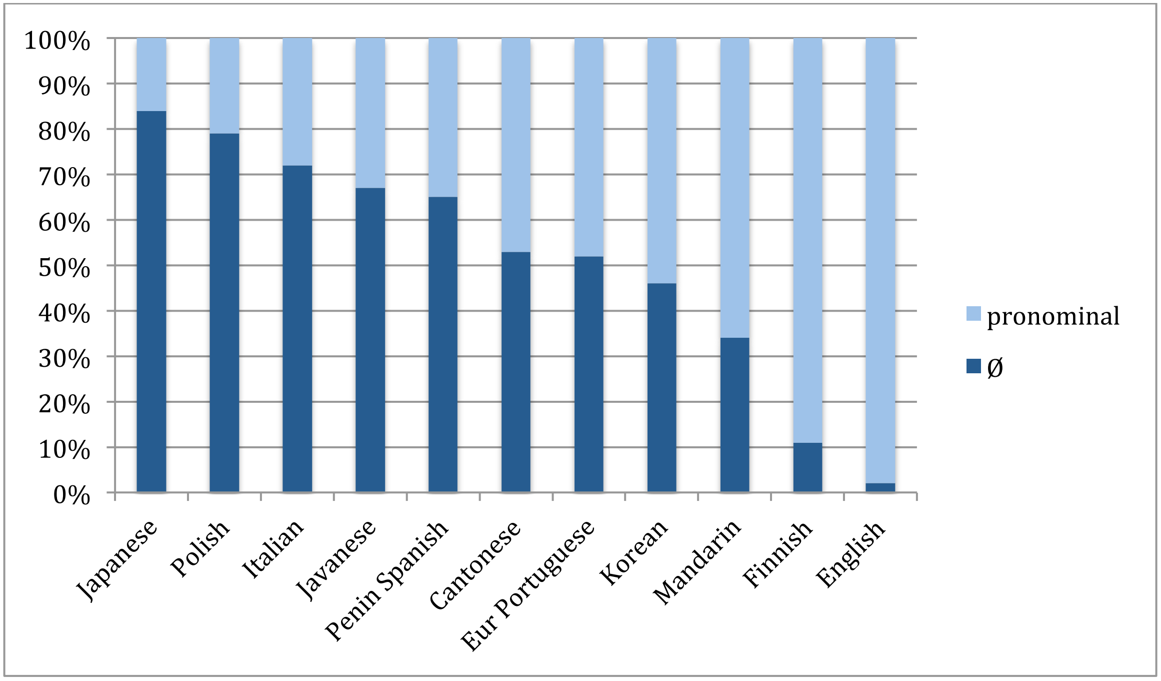
1 Language-internal variation as a window into cross-language variation

Languages are typically classified according to the presence or absence of a given structural feature, in some cases also taking account of that feature's frequency or syntactic distribution (e.g. SOV, case-marking, or ergative languages). A staple of linguistics for decades has been the distinction between null and non-null subject language types (e.g., Rizzi 1982). Functionalist typologists, seeking to identify linguistic universals, and formalist syntacticians, in the tradition of Universal Grammar, largely converge on a classification of language types according to the expression of pronominal subjects. Languages with "obligatory pronouns" or "non-null subject" languages, such as English and French, are opposed to those "in which the normal expression of pronominal subjects is by means of affixes on the verb" or "null-subject" languages, for example, Arabic and Italian (Dryer 2013, Roberts and Holmberg 2010).

Quantitative support for classifying language types has been sought in rates of unexpressed (null) vs. pronominal subjects. Dryer (2013) notes that "if all sentences with pronominal subjects on a couple of pages of text in a language have a pronoun in subject position", the language is taken to be of the "obligatory pronouns in subject position" type. Similarly, a loss of null-subject properties has been inferred from higher subject pronoun rates in Brazilian than in European Portuguese (e.g., Barbosa, Duarte and Kato 2005) and in Dominican than in other varieties of Spanish (Toribio 2000).

Even so, when we consider a sample of languages for which subject expression rates are available, the classification becomes blurred. Figure 1 lists 11 languages by decreasing rate of unexpressed vs. pronominal subjects (restricted to first person singular to control for grammatical person differences). While English is predictably in the rightmost column with approximately 2%, non-expression rates range from 79% to 51% in null-subject languages Polish, Italian, Peninsular Spanish and European Portuguese. As Posio comments, "the actual frequency of subject expression in null-subject languages diverges significantly between languages" (2013: 288).

Figure 1 Rates of 1sg subject expression (Ø vs. pronominal) across different languages



Sources and rates of unexpressed 1sg subjects:

Japanese 84% (Lee and Yonezawa 2008: 738, N=1571), Polish 79% (Chociej 2011: 52, N=536), Italian 72% (Nagy p.c. cf. Nagy et al. 2011, N=224), Javanese 67% (Ewing 2014: 51, N=289), Peninsular Spanish 65% (Posio 2013: 269, N=787), Cantonese 53% (Nagy p.c. cf. Nagy et al. 2011, N=362), European Portuguese 51% (Posio 2013: 269, N=704), Korean 46% (Oh 2007: 466, N=433), Mandarin 34% (Jia and Bayley 2002: 13, N=393), Finnish 11% (Helasvuo 2014: 68, N=1793), English ~2% (Torres Cacoullos and Travis 2014: 22, N=6,600 (estimated)).

Within a single language we also find rate differences. In Spanish, rates of unexpressed 3sg subjects range from 97% (275/285) in Pear Story narratives by Peninsular Spanish speakers (Comajoan 2006: 60) to 73% (122/450) in sociolinguistic interview narratives in Mexico City (Lastra and Butragueño 2015: 43) and 68% (967/1,431) in conversational Colombian Spanish (see Section 9). In English spoken in the USA, rates of unexpressed vs. pronominal 3sg human specific subjects are five times higher in narratives than in conversation (22% (165/748) vs. 4% (153/3,500) respectively, Travis and Lindstrom 2016: 107). Rates of use are thus equivocal, being susceptible to the preponderance or dearth in a data set of some propitious context, which may be fortuitous or due to extra-grammatical, situational considerations, such as register or topic (e.g., Poplack, Zentz and Dion 2012: 250-251; Silva-Corvalán 2001: 163; Travis 2007: 129-131).

Within-language variability is acknowledged to some degree in both functionalist and formalist approaches. Dryer (2013) refers to "normal" subject expression and Roberts and Holmberg (2010: 5), recognizing that unexpressed subjects exist in "non-null subject languages", ascribe to them "special properties that distinguish them from the canonical null subjects". The goal of this paper is to test for the presence of any such "special properties", by comparing quantitative patterns of subject expression in a "non-null subject language"—English—with a well-studied "null subject language"—Spanish.

The status and locus of linguistic universals is much debated. Given the extent of structural diversity across languages (e.g., Evans and Levinson 2009), language universals are said to lie not in structures, which are language-specific, but in the "functions that they perform" (Croft 2001: 60), and in their processes of change, which follow cross-linguistic evolutionary paths (Bybee 2009). Here we contribute to the discussion on universals by attending to the structure of language-internal variability. Proposed functions or processes are operationalized for usage data in the conditioning of variability.

What is the locus of cross-language similarities and differences in actual language use? As the study of language comes to be grounded in quantitative reasoning, more and more linguists recognize the probabilistic structure of grammar proposed by Labov (1969; see also Cedergren and Sankoff 1974 and recently, e.g., Bresnan 2007; Wolk et al. 2013). Probabilistic structure has been important for typology on two fronts, for discovering cross-linguistic tendencies and for determining relatedness among languages. With regard to cross-linguistic tendencies, it has been recognized that these may be manifested in the disfavoring in some languages of structures that are deemed ungrammatical in others (Givón 1979: 22-43). Or as restated, "the same categorical phenomena which are attributed to hard grammatical constraints in some languages continue to show up as statistical preferences in other languages" (Bresnan, Dingare and Manning 2001: 29). As to relatedness among languages, variationist comparative analysis has adapted the comparative method of historical linguistics by incorporating the inherent variability characteristic of speech, for example, in assessing outcomes of language contact (Poplack and Meechan 1998). Variationist comparative analysis considers not merely attestation nor just frequency of some feature but relies on details of co-occurrence and distribution (Poplack and Tagliamonte 1999).

Building on these findings, the proposed dynamic approach of Variationist Typology draws attention to the factors shaping linguistic structure and giving rise to similarities and differences among languages (cf. Chambers 2004). The principle behind Variationist Typology can be summed up as follows.

> ***Variationist Typology***: Cross-linguistic tendencies are manifested in shared aspects of the variable structure internal to each language. Methodologically, similarities and differences across languages are identified through comparisons of intra-linguistic variability. The locus of such comparisons is not only the set of *probabilistic constraints* on the variation but also the delimitation of the *variable context* within which the probabilistic constraints are operative.

In this paper, we apply this approach to the (non-)null subject distinction by probing the structure of language-internal variability in the postulated language types. We rely on comparable corpora of spontaneous conversation from both Spanish and English as the data for analysis, outlined in Section 2. These conversational data allow for a test of whether “VP coordination” is, as widely assumed, a discrete category and a distinguishing feature of English as a non-null subject language. In Section 3, we find that neither of these apply, and thus identify a locus of similarity in what has widely been assumed to be a difference between the two language types. Section 4 presents the first step of the Variationist Typology endeavor, a comparison of the variable context in each of the two languages—where speakers have a choice between an unexpressed and pronominal subject. We discover, again thanks to the conversational data, a prosodic initial-position restriction in English that is absent in Spanish. This, it turns out, is a key locus of difference between the two languages. Sections 5 to 9 take on the second step of Variationist Typology, comparing the probabilistic constraints operative within the variable context. The constraints we test—accessibility, priming, verb class, and tense—are drawn from the large body of work on subject expression cross-linguistically, and our results show remarkable similarities across the two languages, consistent with cross-linguistic patterns. We conclude in Section 10. Precise description and characterization of similarities and differences across languages rests upon both dimensions of the structure of variability in spontaneous speech—the delimitation of the variable context as well as the configuration of probabilistic constraints.

2 Data for comparative analyses

To explore cross-language similarities and differences in spontaneous language use we turn to conversational data.

For English, we use the Santa Barbara Corpus of Spoken American English (SBCSAE) (Du Bois et al. 2000-2005), which consists of 60 recordings (1988-1996), most involving face-to-face conversations between groups of friends, family and acquaintances. In this study, we make use of the 50 transcripts that have tokens of unexpressed 1sg or 3sg human specific subjects.[1] This sub-corpus totals approximately 207,000 words, and includes 88 different speakers of different socio-demographic backgrounds from across the United States.

The Spanish data are drawn from the Corpus of Conversational Colombian Spanish (CCCS, cf., Travis 2005: 9-25). This corpus consists of 30 transcripts of spontaneous conversation between friends and family members (recorded 1997-2004), for a total of approximately 100,000 words, or nine hours of speech. The 38 speakers are primarily middle class, from the city of Cali. 3sg (pronominal and unexpressed) subjects are extracted from all 30 transcripts, and 1sg from the first 21 transcripts, giving similar token counts for each person.

The following examples illustrate the variability in each language. Pronominal subjects have been bolded, and a Ø has been inserted to indicate an unexpressed subject. In the translation of the Spanish, unexpressed subjects are given as pronouns in parentheses.

[1] Ten transcripts (#5, 9, 10, 12, 16, 24, 26, 40, 54, 58) contain no tokens of unexpressed 1sg or 3sg human specific subjects, and were not included in the study.

(1)

Jeff: *So= %,*
... (TSK) (H) he= .. took off for .. Big Bear.
Jill: *... You're kidding.*
Jeff: ***Ø Had** no idea where he was going,*

(SBCSAE 28: 630-633)[2]

(2)

Santi:	*Pero **él nunca va** a la empresa.*	'But **he never goes** to the company.'
Celia:	*.. Claro.*	'.. Of course.'
Santi:	*Entonces,*	'So,
	*ahorita que **Ø necesita** votos,*	now that **(he) needs** votes,
	ahorita sí,	now,
	***Ø llamó** a mecánicos,*	**(he) called** the mechanics,'

(CCCS 01: 848-853)

The data are transcribed prosodically. While the sentence has featured prominently in syntactic analysis, it is not straightforwardly a unit of spoken language (Harvie 1998: 24; Izre'el 2005: 3; Miller 1995: 132). Here we make use of Intonation Units (IUs), segments of speech which are "uttered under a single, coherent intonation contour" (Chafe 1994: 58-60; Du Bois et al. 1993: 47). Each IU appears on a distinct line in the transcription followed by punctuation representing its prosodic contour. For example a period indicates "final" intonation, characterized by a fall to low pitch, and a comma "continuing" intonation, comprising a class of non-final contours (often a slight rise, but also level, or slightly falling pitch) (Du Bois et al. 1993: 53). In (3), a string of IUs linked with commas and ending in a period form what Chafe has described as a prosodic sentence (1994: 139-140). Thus, prosody is relevant to the linking of clauses, as we will discuss below.

(3)

Tom: *(H) And finally,*
when I ran out of money,
uh=,
(H) I wrote home to my family and said,
would you be kind enough to uh,
please send me a passage ho=me.

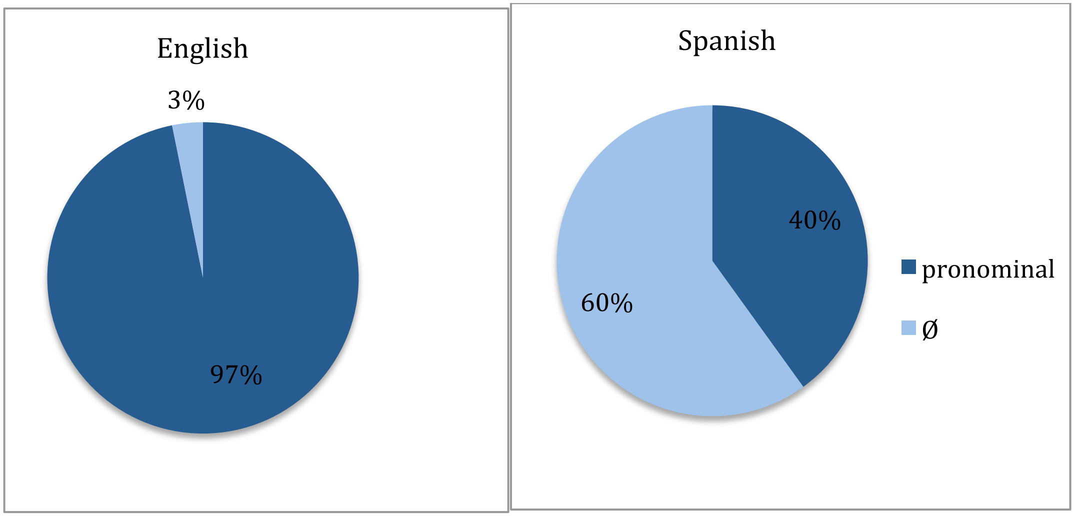
(SBCSAE 32: 508-513)

We focus here on 1sg and 3sg human specific subjects, as these are the most frequent in conversation (Scheibman 2001: 68, 80). For subject expression, in both languages under comparison, the variants are unexpressed and pronominal subjects referring to human participants, be they the interlocutors in the discourse event (first and second person) or persons not present (third person). Lexical subjects (applying only to third person) are set aside, because they may be used to introduce new information (Torres Cacoullos and Travis 2018: Chapter 5).

[2] Examples are reproduced verbatim from the corresponding corpus. Information given in parentheses indicates the corpus, the transcript number and the line numbers of the example. See Appendix 2 for transcription conventions.

As shown in Figure 2, the overall rates of non-expression in English and Spanish are indeed conspicuously different: 60% in Spanish, but approximately 3% in English (counting all tokens of *I* / *he* / *she* / Ø in the 50 transcripts with unexpressed subjects). This has led many to assume that Spanish and English must also be fundamentally different in the patterns of variation between pronominal and unexpressed subjects (e.g., Otheguy, Zentella and Livert 2007: 772, Sorace 2004:144). Given the equivocality of overall rates of use (noted above), in the remainder of the paper we look beyond this, to consider the conditioning factors.

Figure 2 Rates of 1sg and 3sg human specific subject expression
(Ø vs. pronominal):
English (SBCSAE, 329/10,000 (est.)) and Spanish (CCCS, 1,726/2,879)



3 Debunking “VP coordination”

The main observation regarding subject expression from grammars of English is that the subject can be left unexpressed in coordinated contexts (e.g., Biber et al. 1999: 156; Dixon 2005: 22; Quirk et al. 1985: 910). It has been proposed that unexpressed subjects under “VP coordination” are not true null subjects, based on the understanding that verbs under coordination involve a single clause with two predications, rather than two clauses with a null subject in the second (cf. Haspelmath 2004: 31; Huddleston and Pullum 2002: 238). Alternatively, scholars have appealed to the rule of Conjunction Reduction, according to which two sentences are reduced to one under coordination (e.g., Akmajian and Heny 1980: 261-262). Such claims have been made, however, in the absence of an operational definition of VP coordination.

Thus, a first question is whether English has genuine null subjects. To answer this, we look beyond theory-internal convictions and test empirically whether VP coordination is a discrete category, and whether it is a distinguishing feature of English as a non-null subject language.

The following, excerpted from (3) above, illustrates what would be considered a canonical case of VP coordination.

(4)

Tom: I wrote home to my family and **Ø said**,

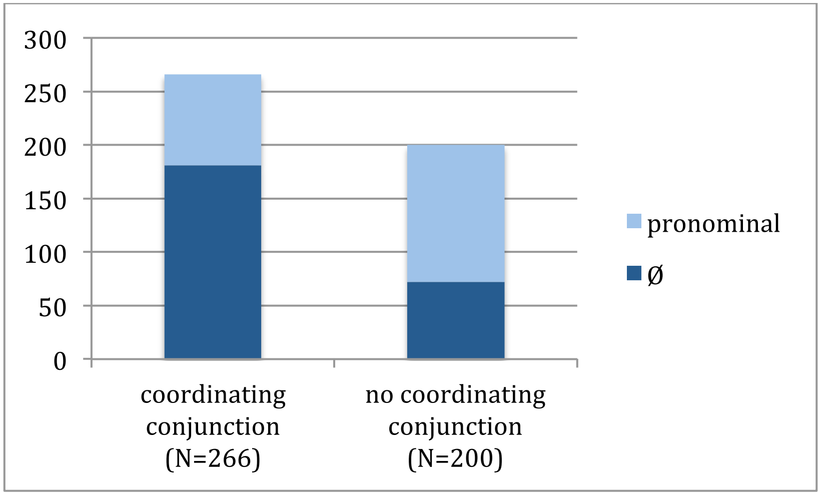
(SBCSAE 32: 511)

Here we clearly have tight linking between the two verbs which occur in the same prosodic unit and are coordinated via the conjunction *and*. However, notions such as coordination and conjunction remain nebulous (e.g., Matthews 1981: 217). VP coordination is generally applied, first, to the occurrence of an unexpressed subject that is coreferential with a preceding subject (Quirk et al. 1985: 948). Beyond this, two sets of features that emerge from the literature are conjunction presence and prosodic relation. Coordinate clauses/verbs with coreferential subjects may be connected with or without a conjunction, in what has been termed syndetic vs. asyndetic coordination (Biber et al. 1999: 156; Haspelmath 2004: 4), and/or they may be prosodically connected (e.g. Mithun 1988: 332; Quirk et al. 1985: 948). How does each of these features hold up empirically?

We begin with *syntactic* linking, considering the presence of a conjunction. The coordinating conjunctions found in the English data are *and, but,* and *or* (Quirk et al. 1985: 910), and these in combination with an adverb, the most frequent combination being *and then*. We compare coreferential clauses with a coordinating conjunction vs. with no such conjunction in Figure 3, and coreferential clauses with *and* vs. with other coordinating conjunctions in Figure 4. The height of the columns in these figures indicates the number of tokens, and the darker shading the proportion of unexpressed subjects. We find that the rate of non-expression in coreferential contexts is notably higher when a conjunction is present (syndetic coordination) than when it is not (asyndetic coordination). As seen in Figure 3, the difference is 68% (181/266) vs. 36% (72/200) (*p* = 0.0001, Fisher's exact test), in a sample of the data with an artificially high overall rate due to sampling method (see Section 5.1). Within the syndetic contexts, Figure 4 illustrates that coordination with *and* is by far the most frequent (occurring over five times as often as all other coordinating conjunctions combined—*but, or,* and *and/but* + ADV). It also appreciably has the highest proportion of unexpressed subjects, 74% (169/228) vs. just 34% (13/38) with other conjunctions (*p* = 0.0001).

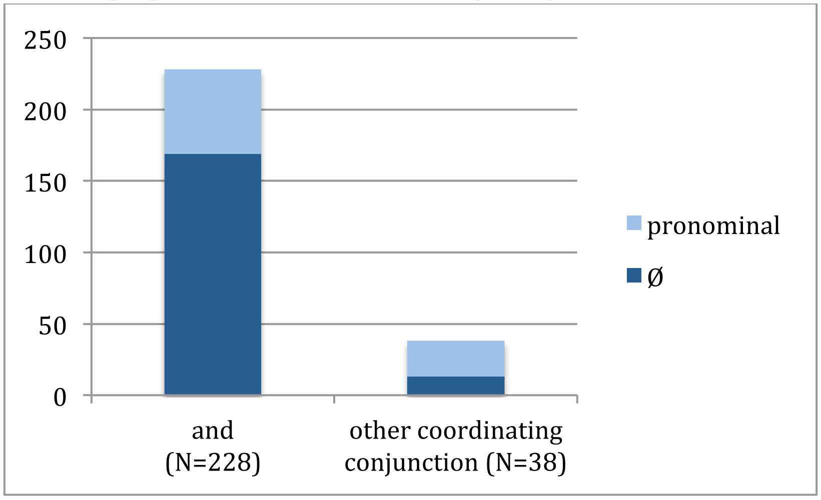
Thus, in operationalizing coordination, among the various forms routinely included in the literature, usage data obliges us to single out coordination of coreferential-subject clauses with *and* as distinct both from asyndetic coordination and from syndetic coordination with conjunctions other than *and*. It is *and*-coordination that realizes syntactic linking.

Figure 3 Syntactic linking – Number of occurrences of coreferential contexts according to the presence of a conjunction
and proportion of Ø (darker shading): English



Overall rate Ø = 54%

Figure 4 Syntactic linking – Number of occurrences of coreferential contexts with *and* vs. other coordinating conjunctions
and proportion of Ø (darker shading): English



Overall rate Ø = 68%

The second set of features concerns *prosodic* linking, which is integral to connecting elements in speech. To test for the role of prosody in coordination, we measure prosodic connectedness across coreferential clauses by relying on the boundary and contour of the prosodic unit. Clauses that occur either in the same IU or in adjacent IUs where the first has continuing intonation contour are considered to be prosodically linked (cf. Chafe (1988: 10) on English and Mithun (1988: 332) cross-linguistically). We consider this prosodic linking together with syntactic linking (with *and*). For two clauses/verbs with coreferential subjects, there are four possible linking configurations, listed below with corresponding examples following.

1) maximal: linked both syntactically and prosodically (in the same IU, (5), or across IUs separated by a comma, (6));
2) intermediate: linked only syntactically (7);
3) intermediate: linked only prosodically (8)
4) none: no prosodic nor syntactic link (9).

Note that in each configuration there occur both unexpressed subjects (illustrated in (a) in each pair of examples) and pronominal subjects (illustrated in (b)).

(5) ✓ syntactic linking (*and*)
✓ prosodic linking (same Intonation Unit)
a. *I wrote home to my family and **Ø said**,* (SBCSAE 32:511)
b. *... and then I go and **I talk** to him.* (SBCSAE 21:1098)

(6) ✓ syntactic linking (*and*)
✓ prosodic linking (continuing (comma) intonation contour)
a. *... Dad called him**,***
*and **Ø told** him he had to.* (SBCSAE 31:363-364)
b. *so he came**,***
*(H) and **he stood** opposite me,* (SBCSAE 55: 161-162)

(7) ✓ syntactic linking (*and*)
✗ prosodic linking (e.g., final intonation contour)
a. *and he ran them off**.***
*... And **Ø saved** their lives.* (SBCSAE 30: 572-573)
b. *he's a broker**.***
*... And **he buys** hay,* (SBCSAE 56: 803-804)

(8) ✗ syntactic linking (no *and*)
✓ prosodic linking (continuing (comma) intonation contour)
a. *A=nd then I worked for a rancher over there for a while**,***
*... **Ø followed** the rodeos for a while,* (SBCSAE 32: 1587-1588)
b. *That's what I did all day today**,***
***I had** ... three or four different kids come up,* (SBCSAE 43: 156-157)

(9) ✗ syntactic linking (no *and*)
✗ prosodic linking (e.g., final intonation contour)
a. *... And yesterday was the first day she used it**.***
*(H) **Ø Put** a bunch of stuff in it to read,* (SBCSAE 43: 34-35)
b. *.. I do the hard labor**.***
***I build** barns and,* (SBCSAE 56: 84-85)

Looking for verification that transcends theoretical orientations, and recognizing the circularity of the argument that the presence of the subject pronoun itself annuls VP coordination, we make a prediction that can be quantitatively falsified. If coordination is a discrete category of false null subjects, and the second verb is not a genuine clause, we expect categorical behavior —100% unexpressed subjects—in at least the configuration with the tightest linking. But in fact, none of the configurations is exclusive of either variant, as exemplified in (5) through (9), and as demonstrated in Figure 5, which gives the rates of subject expression across the four configurations. Instead of categorical behavior we find gradience: the tighter the link—*prosodic* and *syntactic*—between the target and the preceding clause coreferential subject, the higher the rate of unexpressed vs. pronominal subjects.

If, further, coordination were a special property of English as a non-null subject language, we would expect to observe differences with a bona fide null subject language. Figure 6 shows that this is not the case, since linking in Spanish also has a graded effect on subject expression. In both languages, the rate of unexpressed subjects is highest with both prosodic and syntactic linking, lower with one or the other kind of structural linking, and lowest in the absence of linking.

Figure 5 Rates of Ø according to prosodic and syntactic link to preceding clause (coreferential contexts): English

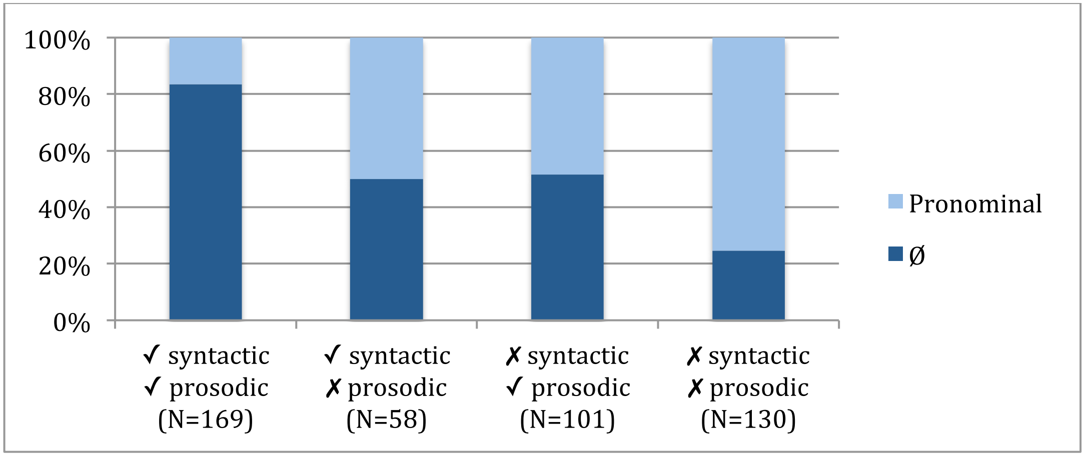
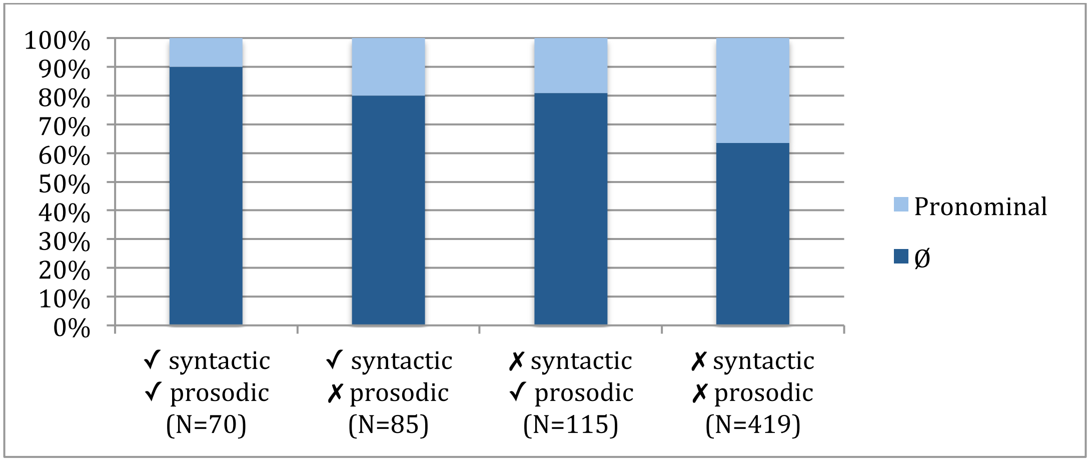


Figure 6 Rates of Ø according to prosodic and syntactic link to preceding clause (coreferential contexts): Spanish



In sum, once we operationalize coordination and test it against the data, contrary to what has been widely assumed, so-called VP Coordination in English shows neither categorical nor special behavior. These facts compel us to conclude that VP Coordination must be abandoned both as a discrete category and as a property of English that sets it apart as a non-null subject language.[3]

Moving forward, these facts also dictate that unexpressed subjects under coordination—including maximal linking, with *and* along with prosodic connection—be included in the quantitative analysis of variation as part of the variable context in English. Further, these results bring out linking to the preceding subject as a factor conditioning unexpressed subjects that should be tested in both languages (see Section 6).

We now implement Variationist Typology, beginning with delimitation of the variable context before considering the probabilistic constraints.

4 Differences in the variable context

While increased attention to probabilistic grammars has highlighted constraints on variable use, including direction and strength of effect, we suggest that a dimension of probabilistic grammar that deserves to be brought to the fore in the enterprise of cross-linguistic comparisons is the *variable context*. The variable context, the broadest domain in which speakers have a choice between variants, must be accurately delimited in order to discover the constraints in operation where speakers have a choice (Labov 2005:7). The envelope of variation is thus the first consideration for Variationist Typology, as a key locus of comparison. Is the variable context—here, the set of linguistic contexts where both pronominal and unexpressed subjects occur— the same or distinct in the proposed different language types?

4.1 The prosodic initial-position constraint in English

As seen in Section 3, coreferential-subject clauses conjoined with a coordinating conjunction are variable as to subject expression. However, outside of coordinate clauses, unexpressed subjects in English occur solely in prosodic-initial (IU-initial) position (Torres Cacoullos and Travis 2014: 25-26). In (10), for example, the IU-initial pronoun in *he doesn't* and unexpressed subject in *Ø has no idea* occur in the variable context; the IU-medial *he* in *he's gay* does not. This initial-position restriction means that, outside of coreferential-subject verbs with a coordinating conjunction, the variable context for subject pronoun expression in English is restricted to declarative main clause prosodic-initial-position verbs.

[3] Thanks to Joan Maling for pointing out that case marking counters VP-coordination, for example, in Icelandic, which is not a null-subject language and the null subject in the second conjunct can be a different morphological case than the overt subject in the first conjunct. In the following example from Rögnvaldsson (1982: 559), the subject of the 2nd verb would be dative if overt.

Þeir	sjá	stúlkuna	og	[e] finnst	hún	álitleg.
they-NOM	see	the.girl-ACC	and	find	she-N	attractive

‘They see the girl and find her attractive’

(10)

Cam: .. ***He$_1$ doe=sn't know*** *that* ***he$_2$'s*** *gay?*
Lajuan: .. *(THROAT) Hm-mm.*
Cam: .. *(THROAT)*
Lajuan: .. ***Ø$_1$ Has*** *no idea.*

(SBCSAE 44: 250-253)

The restriction to initial position for English null subjects has been interpreted as a syntactic constraint. Under the formal syntactic view, null subjects have been considered a root, or main clause, phenomenon (e.g., Haegeman 2013), and quantative data has shown a favoring of null subjects in clause-initial position in English (and French) (Harvie 1998: 21; Leroux and Jarmasz 2005: 7). However, the prosodically-transcribed speech data here allow us to establish that the initial-position restriction is in fact prosodic, and that the lack of variability in subordinate clauses and interrogatives is bound up with prosody. In subordinate clauses and interrogatives, subject pronouns simply do not occur in IU-initial position (in a sample of 500 tokens made up of the first 10 in each of the 50 conversations under study here; N subordinate = 96, N interrogative = 16).[4] Main clause subject pronouns, on the other hand, do occur in non-initial position, for example, preceded on the IU by *well*, *oh, of course* (approximately one quarter of the time, 99/388). This allows us to tease apart sentence or clause type and prosodic position: if the constraint were a purely syntactic one, then unexpressed subjects in declarative main clauses should also occur in non-initial position. But in fact outside of coreferential-subject clauses conjoined with a coordinating conjunction, unexpressed subjects never occur in non-IU-initial position. Thus, the restriction is prosodic rather than syntactic, consistent with previous phonological accounts of unexpressed subjects in English as "left-edge deletion" (Weir 2012; cf. Napoli 1982; Quirk et al. 1985: 896; Sigurðsson and Maling 2010).

As well as non-IU initial subjects (outside coordination), also excluded from the English variable context are pronouns occurring with contracted auxiliary forms (e.g. *he's, I'm*) and particular formulaic expressions, for which we treat the pronoun as integral to the formula (or "chunk"). This is the case for all instances of *I mean*, and for discourse marker uses of *I guess, I think* and other 1sg expressions, defined as instances occurring on their own in an IU or as parentheticals (Travis and Torres Cacoullos 2014: 366). Also invariable are quotatives, such as *I said*, *she goes* (except under coordination).

4.2 A narrower variable context in English vs. Spanish

The variable context in Spanish is much less restricted than that of English, and this represents a sharp difference between the languages. The major factor is position in the prosodic unit: Spanish shows robust variability in both IU-initial and non-IU initial position, and correspondingly, in both main and subordinate clauses.[5]

[4] There is only one IU-initial subject pronoun in a subordinate clause which, note, is flagged through pausing, lengthening and an in-breath.

Doris: *and ask if= uh,*
... (H) ***he****'s still speaking to me,* (SBCSAE 11:26-27)

[5] The IU-initial position restriction of unexpressed subjects in English is not even a statistical preference in Spanish, where unexpressed subjects are not favored in IU-initial over non-IU initial position (55%, 972/1,761 and 66%, 541/823, respectively).

Furthermore, while in English we identified and excluded discourse formulae such as *I mean*, in the Spanish data here, the constructions (*yo) creo* 'I believe/think' and (*yo) no sé* 'I don't know' strongly favor expression (Section 8.2), but nevertheless remain variable, and thus are included (Travis and Torres Cacoullos 2012: 738-742).

Finally, although contrast (or emphasis) has been widely surmised to be a general function served by subject pronouns in null-subject languages (e.g., Chafe 1976: 37; Haegeman 1994: 217; Payne 1997: 43), operationalizations of this notion have revealed that it offers a very limited account of subject pronoun expression in Spanish. Contrastive constructions, for example with converse predicates, do show relatively high subject pronoun rates, but they constitute a small proportion of subject pronouns (Travis and Torres Cacoullos 2012: 714-724). And while particular cases of pronouns deemed to be contrastive have been excluded from some previous studies of Spanish (e.g., Silva-Corvalán 2003: 850), pronouns are variable in contrastive contexts (Amaral and Schwenter 2005; Otheguy and Zentella 2012: 233). Thus only morphologically marked tokens of emphasis are excluded, here the so-called emphatic constructions in which the pronoun is followed in the same IU by *mismo* 'same' (e.g. *yo mismo*, 'I myself') or *sí* 'yes' (e.g. *yo sí* 'I yes'). These are, however, notably rare. Out of over 1,500 tokens of 1sg and 3sg pronouns in the data, the former does not occur, and the latter occurs only ten times.

Figure 7 summarizes the set of environments constituting the variable context in each language.[6] Note that, while in English the proportion of subject pronouns occurring within the variable context is just 42% (based on the sample of 500 pronouns referred to above), in Spanish 97% of the tokens occur within the variable context. Our first finding, then, in implementing Variationist Typology, is that one locus of difference between English and Spanish is the much more limited variable context in English, driven primarily by the prosodic-initial position restriction. This finding lays bare a specifiable manifestation of their classification as null vs. non-null subject language types.

[6] We do not count subject relatives in either language (as the subject is expressed with a relative pronoun), and for Spanish, postverbal pronouns (N=44) since they are sensitive to different conditioning factors from unexpressed and preverbal subjects (Silva-Corvalán 1982: 113). *Wh*-interrogatives which are either unexpressed or expressed in postverbal position in this variety are outside the variable context.

Figure 7 Variable context for subject pronoun expression: English vs. Spanish

	English **(N=500)**	**Spanish** **(N=2,879)**
Variable context	**42%**	**97%**
Coordination	Coordinated (coreferential-subject clauses with coordinating conjunction)	
Prosody / Clause type	Prosodic-initial position -Main clause declaratives	Prosodic-initial & medial position - Main clause declaratives - Subordinate clauses
Pronoun categorically present or absent	**58%**	**3%**
	Prosodically medial position (outside coordinated contexts) (39%)[7] - Main clause declaratives - Subordinate clauses - Interrogatives Contracted forms (13%) Discourse markers (5%) Quotatives (outside coordinated contexts) (1%)	Morphologically marked emphatic construction “*yo / él / ella* + *sí* + V” (lit. ‘I / he / she + yes + V’) (0.5%) Spanish: *wh*-interrogatives (2.5%)

5 Similarities in the probabilistic constraints

Having delimited the envelope of variation, we can now set aside non-variable contexts where one or the other subject form is categorically present and move on to the second step of Variationist Typology: to compare the internal structure of the variability within the variable context. This is accomplished through discovery of the *linguistic* conditioning of variant selection in each language, that is, probabilistic statements about the co-occurrence of variant forms and elements of the linguistic context in which they appear. We have seen that Spanish and English differ in their envelopes of variation. One might predict that the probabilistic constraints will also be different across the two languages. Are they?

[7] Instances with more than one feature outside the variable context were classified following the order of exclusions indicated in Figure 7. Thus, the percentages given for contractions, discourse markers and quotatives under-represent the N of their occurrences.

5.1 Principled sampling of a rare variant

Essential for the analysis are comparable datasets in each language, datasets that are circumscribed to the respective envelope of variation and in which both of the variants are robustly represented in a principled way.

For Spanish, in accordance with the variable context just described, we excluded interrogatives, and the “emphatic” pronoun+*sí* construction, leaving a rate of 59% unexpressed (1,659/2,802).

For English, given the very low rate of non-expression, we took a principled sample of pronouns, thus creating an artificially higher relative frequency of the unexpressed variant for the statistical analysis (cf. Harvie 1998: 18; Leroux and Jarmasz 2005: 3; Torres Cacoullos and Travis 2014: 23; Travis and Lindstrom 2016: 109). For each unexpressed subject we extracted the closest preceding and following subject pronoun of the same grammatical person produced by the same speaker, and falling *within the variable context*. With the large proportion of the data that non-variable contexts account for (Figure 7), the immediately preceding or following pronoun was eligible for the sample in less than one half of the instances.

By way of illustration, in the following example, of the four 3sg subject pronouns, only one falls within the variable context and is available for extraction, namely that in line 3; none of those in lines 6, 7 and 10 occur in IU-initial position or are conjoined via a coordinating conjunction with a coreferential-subject clause. (Further, that in line 6 also occurs in an interrogative, and is produced by a speaker distinct from that who produced the unexpressed, and that in line 7 also occurs in a subordinate clause).[8]

(11)			pronoun in variable context
1.	Corinna:	*.. Well it took em like tw- --*	
2.		*over twenty years to catch this guy.*	
3.		***He went** and Ø shot his entire f=amily.*	✓
4.		*... And then Ø disappeared.*	
5.	Patrick:	*... Well?*	
6.		*... So why did **he shoot** his family.*	✗
7.	Corinna:	*... Cause **he was** afraid that their morality was going downhill.*	✗
8.	Patrick:	*... Well,*	
9.		*... yeah[=]?*	
10.	Corinna:	*[So] **he shot** his mother,*	✗
11.		*his kids and his wife.*	
12.		*... Oh and his daughter.*	

(SBCSAE 45: 670-681)

[8] When two or more unexpressed tokens occurred sequentially, as in this example, we extracted the two closest preceding and following tokens from the same speaker. When there were no available tokens from the same speaker, we extracted from another speaker who participated in the conversation and who produced at least one unexpressed token.

This resulted in an English data set of 878 observations, and a modified (artificial) rate of unexpressed subjects of 38% (329/878).[9] The point is to examine the relative frequency of the variants in linguistic sub-contexts, in order to discern the effect that those sub-contexts have on speaker choice to express the subject or not.

5.2 Probabilistic constraints

Here, building on the large body of work on subject expression cross-linguistically, we consider the candidate constraints of accessibility, priming, verb class, and tense.

We conduct separate logistic regression analyses on the two datasets, with the independent variables (predictors) configured identically so that differences and similarities in direction of effect are readily visible.[10] Table 1 shows the results of these analyses, with English and Spanish juxtaposed. For each predictor, listed on the left, are the elements of the linguistic context that constitute the levels of the predictor. Variable-rule analysis determines the predictors that together account for the largest amount of variation, in terms of stepwise increase of log likelihood, such that the addition of any of the remaining predictors does not significantly increase the fit to the model (Sankoff 1988). (For Generalized linear mixed effects models see Appendix 1.)

The first column in Table 1 gives, for each level, the relative probability value for an unexpressed subject such that the closer to 1 the more an unexpressed subject is favored in that context, and the closer to 0, the stronger the disfavoring effect, or, conversely, the favoring of the pronominal variant. The following columns give the rate of unexpressed subjects in that context (% Ø), the number of tokens in that context (N) and the proportion of the data that context accounts for (% data).

Selected as having a significant effect (indicated in bold) are Linking to the preceding subject (a reconfiguration of accessibility), Coreferential Subject Priming, and Verb class, while Tense is not selected as significant in either language. We see here that, *in both languages*, unexpressed subjects are most favored with prosodic and/or syntactic linking to the preceding coreferential subject, when the previous coreferential subject was also unexpressed, and with dynamic over stative verbs. Our second finding towards Variationist Typology, then, is that constraints across Spanish and English are shared. Let us now consider each of these constraints.

[9] The English sample rate of Ø is not exactly one third since after first extracting two pronouns (meeting all other restrictions to the variable context) for each Ø, we excluded all non-IU-initial pronouns.

[10] To compare different languages (or varieties of a single language) we opt for separate analyses of each dataset and juxtaposition of the models rather than a single model that includes the element of comparison (e.g., language, time period) as a fixed effect and interaction term, in order to highlight (dis)similarities in direction of effect.

Table 1: Two independent logistic regression analyses of factors contributing to the choice of Ø vs. pronominal 1sg and 3sg human specific subjects, in conversational English (SBCSAE) and Spanish (CCCS) (using variable rule analysis).*

	English				**Spanish**			
	N=329/878; Input: .34				1,659/2,802; Input: .60			
	(Overall rate: 38%)				(Overall rate: 59%)			
	Prob	**% Ø**	**N**	**% data**	**Prob**	**% Ø**	**N**	**% data**
Linking to preceding subject (accessibility)								
Semantic (coreferential) + prosodic and/or syntactic link	.78	68%	328	38%	.70	79%	458	17%
Semantic (coreferential) only	.41	27%	192	22%	.53	63%	700	26%
No link (non-coreferential)	.27	15%	344	40%	.43	52%	1588	58%
Range	*51*				*27*			
Coreferential Subject Priming								
Previous mention as unexpressed	.63	57%	115	13%	.63	73%	841	32%
Outside priming environment**	.45	23%	299	35%	.48	55%	1138	44%
Previous mention as pronoun	.50	42%	448	52%	.37	46%	637	24%
Range	*19*				*26*			
Verb class								
Dynamic	.54	45%	573	66%	.57	67%	1761	63%
Stative	.45	24%	179	21%	.42	53%	685	25%
Cognition	.41	25%	114	13%	.31	35%	335	12%
Range	*20*				*26*			
Tense								
Past / Preterit		40%	651	76%		66%	667	27%
Imperfect						58%	290	12%
Present		33%	200	24%		55%	1490	61%

* Probability values given only for significant predictors (p < .05)[11]

** Outside priming environment: for both languages, previous mentions as full NPs and relative pronouns, previous or target in quoted speech, and previous produced by interlocutor; for English, previous mentions beyond 5 clauses; for Spanish previous mentions beyond 10 clauses and previous post-verbal subjects.

6 Accessibility as linking to preceding subject

A widely reported constraint on subject expression, and a candidate universal conditioning factor, is what has been termed “accessibility” (Ariel 1994: 2630; Givón 1983: 17), “activation” (Chafe 1994: 75) or “recoverability” (e.g. Haegeman 2013: 89; Weir 2012:

[11] Variable-rule analysis *p*-values are not reported because they are a measurement of how likely the change of likelihood is due to chance when a factor group is added to the model and so are meaningful in the context of the entire series of “stepping up” and “stepping down” runs.

116). The observation has been that the greater the “accessibility”, “activation”, or “recoverability” of the subject referent, the greater the likelihood that the subject is unexpressed. While there are different paths of accessibility, one prominent measure is distance from the preceding mention in terms of clauses or intervening human subjects (Givón 1983: 12-14; Travis and Torres Cacoullos 2012: 720-723). In quantitative work on subject expression it has generally been configured as coreferentiality with the immediately preceding clause subject. Coreferentiality conditions variable subject expression across a range of languages. Same reference has been found to favor unexpressed subjects and conversely a switch in reference to favor pronominal expression in Arabic (Owens, Dodsworth and Kohn 2013: 263), Australian Sign Language (McKee et al. 2011: 388), Persian (Haeri 1989), Bislama (a Vanuatan creole) and Tamambo (an indigenous language of Vanuatu, Meyerhoff 2009: 308), as well as Spanish (e.g., Cameron 1994: 32; Silva-Corvalán 2001: 154) and English (Harvie 1998: 21; Leroux and Jarmasz 2005: 7; Torres Cacoullos and Travis 2014: 24; Travis and Lindstrom 2016: 112).

Here the measure of coreferentiality with the preceding clause subject is refined, following from what we saw in Section 3, by bringing in both syntactic and prosodic linking between clauses. Thus, our proposal is that “linkage to antecedents” (Levinson 1987: 381) or “conjoinability” of clauses (Li and Thompson 1979: 330) be treated as a *combination of semantic and structural features*. Accordingly, the semantic link of coreferentiality is broken into two categories based on the presence or absence of syntactic and/or prosodic linking. As can be seen in Table 1, unexpressed subjects are most favored precisely when the preceding clause subject is coreferential and there is also a prosodic and/or syntactic link between the target and preceding clause (as in examples (5) – (8) above), less so in coreferential contexts when there is no such structural linking, as in (9), and least in noncoreferential contexts, as in (10).[12]

Note that structural linking with the preceding clause impacts unexpressed subjects more than mere coreferentiality with the preceding subject does. In both Spanish and English there is a bigger difference between prosodically and/or syntactically linked coreferential contexts vs. coreferential contexts with no linking than there is between merely coreferential vs. non-coreferential (switch reference) contexts. This result suggests that it will be profitable to rethink accessibility in discourse as a composite of semantic and structural features—both syntactic and prosodic (Torres Cacoullos and Travis 2018, Chapter 5).[13]

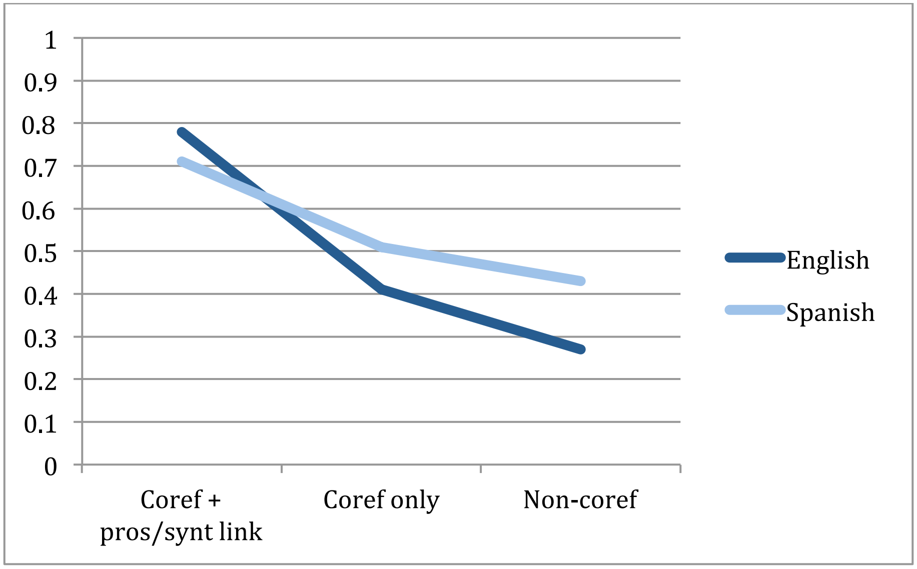
One key difference across the two languages concerns the relative strength of the effects. An indication of relative strength is the Range for each predictor, calculated as the difference between the most and least favoring probability values. In English, Linking

[12] Included in coreferential contexts lacking structural linking are previous mentions produced by an interlocutor (English N=62; Spanish, N=300). The rate of Ø in this context changes little if we exclude such tokens, 25% (32/130) in English, and 63% (253/400) in Spanish.

[13] It might be argued that prosodic and syntactic linking are masking semantic effects. Cross-tabulation of linking and temporal sequencing (see Section 8.1) confirms that a greater proportion of structurally linked tokens (as opposed to merely coreferential) are indeed temporally related, but the rate of Ø remains higher with structural linking than without even in non-temporally related contexts (78% (130/166) vs. 62% (272/442), in the Spanish data).

(accessibility) clearly is a stronger constraint than either Priming or Verb class (displaying a Range that is more than double that of the other significant predictors), while in Spanish each of these predictors is of similar magnitude. The strength of Linking in English is seen in the steeper drop in the probability of an unexpressed subject from the relative favoring in structurally linked contexts to the strongly disfavoring in non-coreferential contexts (illustrated in Figure 8). The quantitative comparative analysis applied in the Varitionist Typological approach allows us to discern a locus of difference in strength of effect, while revealing that the trend, or direction of effect, in the two languages is the same.

Figure 8 Probability of Ø according to linking with preceding clause subject (Table 1)



7 Coreferential Subject Priming

Structural priming is the repetition of the same syntactic structure across clauses without pragmatic motivations, what Labov has termed a “mechanical effect” (1994: 547–568; cf. Bock and Griffin 2000; Gries 2005; Szmrecsanyi 2005). For subject expression, a tendency for unexpressed subjects to follow unexpressed subjects and pronouns to follow pronouns has been observed across a range of languages (e.g. Australian Sign Language (McKee et al. 2011: 388), Bislama and Tamambo (Meyerhoff 2009: 308), Spanish (e.g., Cameron and Flores-Ferrán 2003; Travis 2007: 120-122) and English (Torres Cacoullos and Travis 2014: 29-30; Travis and Lindstrom 2016: 115)).

Coreferential subject priming—repetition of the form of the previous coreferential subject—is illustrated in the following examples, where targets and primes are bolded. Unexpressed to unexpressed priming in English is shown in (12) and (13), and pronominal to pronominal priming in Spanish in (14) and (15). As seen here, priming occurs in linked ((12), (14)) as well as in non-coreferential contexts ((13), (15)).

(12)

Darlene: *.. (H) He invented it?*
.. And ***Ø started*** *his own business,*
and ***Ø sold*** *it to other doctors,*

(SBCSAE 52: 1164-1166)

(13)

Fred: *... (H) I look at my bank sta- .. bank statements.*
And,
Wess: *Mhm.*
Fred: *...* ***Ø look*** *through my checks.*
When they come in.
.. And ***Ø make sure*** *that it's fine.*

(SBCSAE 59: 749-754)

(14)

María:	*..* ***Yo me montaba*** *en bus,*	'**I would get on** a bus,
	yo no cogía *taxi para --*	**I wouldn't take** a taxi to --
	pa' mi mercado.	to do my shopping.'

(CCCS 13: 472-474)

(15)

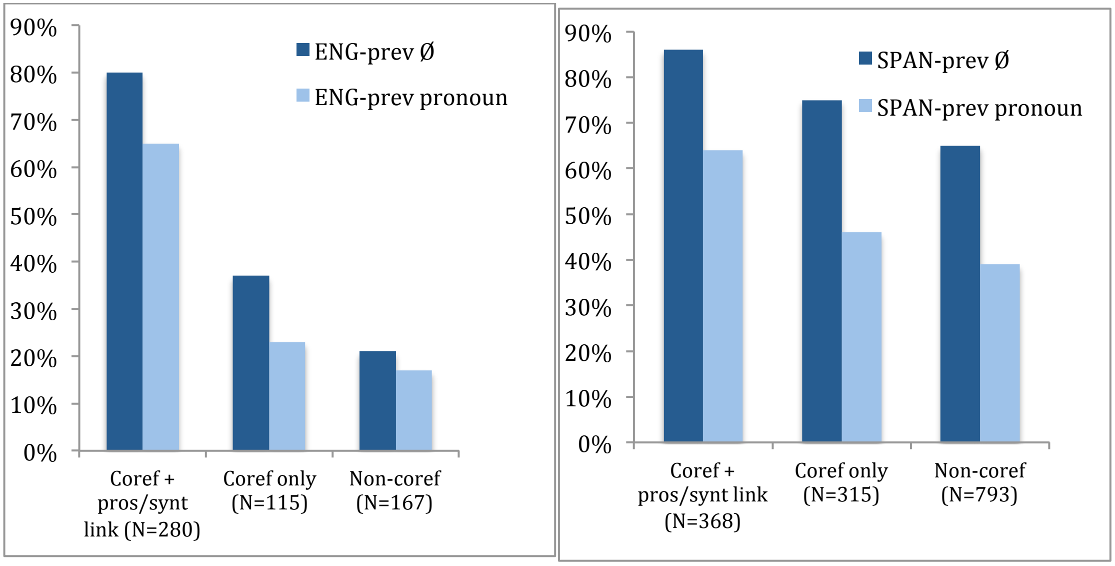
Ángela:	*no sé si fue que* ***él le estaba haciendo*** *muy rápido,*	'I don't know if it was that **he was doing it** very quickly,
	Y el computador pues está como muy lento,	And the computer is kind of very slow,
	Y ***él como que le daba*** *muchas órdenes al mismo tiempo,*	And **he kind of gave it** a lot of commands at the same time,'

(CCCS 29: 467-469)

Priming and accessibility are independent effects. Figure 9 displays the rate of unexpressed subjects according to the realization of the previous coreferential subject (as either pronominal or unexpressed) by linking to the preceding clause. In both English (on the left) and Spanish (on the right), the darker column (previous unexpressed) is taller than the lighter one (previous pronoun) demonstrating that the priming effect applies in each of the three linking contexts. There is nevertheless an interplay between the effects (Cameron 1994: 140; Travis, Torres Cacoullos and Kidd 2017: 289). In particular, in English, coreferentiality and unexpressed-to-unexpressed priming are bound together (see Torres Cacoullos and Travis 2014: 29-30).[14]

[14] In Figure 9, English rates of unexpressed subjects when the previous mention is unexpressed vs. pronominal, in linked, coreferential only, and non-coreferential contexts, respectively are: 80% (53/66) vs. 65% (140/214) (p = 0.02), 37% (7/19) vs. 23% (22/96) (p = 0.25, n.s.), and 21% (6/29) vs. 17% (24/138) (p = 0.79, n.s.). The corresponding figures for Spanish are 86% (190/220) vs. 64% (95/148), 75% (134/179) vs. 46% (62/136), and 65% (288/441) vs. 39% (137/352) (p < 0.0001 in all cases by Fisher's exact test).

Figure 9 Rate of Ø according to priming and linking: English (left panel) and Spanish (right panel)



8 Verb class

Verbal semantics and lexical aspect play a role in Spanish, with cognition verbs favoring 1sg expressed subject *yo* 'I' (e.g., Bentivoglio 1987: 48-53; Enríquez 1984: 152, 235-245) and dynamic predicates favoring unexpressed subjects more than do statives, especially among frequent verbs (Claes 2011: 205, Erker and Guy 2012: 541-542). We therefore coded all verbs into three classes: cognition, dynamic (or "external action"), and stative. Both languages exhibit an identical effect for dynamic verbs, which favor non-expression, while cognition verbs strongly favor expression in Spanish, but not in English. There is no single explanation for this behavior; rather we must appeal, on the one hand, to an effect of temporal sequencing, and on the other, to language-specific particular constructions.

8.1 Dynamic verbs and temporal sequencing

Coded as dynamic are a range of verbs of action (e.g., *put, do, give*), motion *(go, come*), speech (*say, tell*), and perception (*see, look*). The effect of dynamic verbs found in both languages is not a verb class effect per se, but is tied to a favoring of unexpressed subjects in clauses that are temporally related to the preceding clause, or conversely, a disfavoring in non-temporally related contexts. Such an effect for temporal sequencing was first observed for Brazilian Portuguese, under what was termed "discourse connectedness" (Paredes Silva 1993: 43).

Temporally related clauses are defined as consecutive main clauses with coreferential subjects that refer to events or situations which are temporally sequential or simultaneous (Travis and Lindstrom 2016: 116-119; cf. Labov and Waletzky 1997 [1967]: 12-13). Examples (16) and (17) illustrate temporal sequentiality, and examples (18) and (19) clauses referring to simultaneous events. Non-temporally related clauses include those where at least one of the verbs in the sequence is a stative verb, such as 'have' in (1) (*Ø had no idea*); occurs with habitual aspect (as in (7b) (*he buys hay*) and (9b) (*I build barns*), unless the events are habitually sequential or simultaneous, as in (19)); or is a repetition of, or elaboration on, the previous verb, as in (20). Temporal relationship is not applicable when the previous coreferential subject was produced by the interlocutor or was part of quoted speech. Subordinate clauses, also, are generally not relevant to temporal sequentiality (Labov and Waletzky 1997 [1967]: 14).[15]

(16)

Jo: *... So I said that's not ... as per our agreement,*
he said yes it is,
he handed *me my check,*
Ø rolled up *his window,*
and ***Ø drove*** *off.*

(SBCSAE 53: 289-293)

(17)

Celia	***Yo la pongo*** *encima de la mesa?*	'**I put it** on top of the table?
	... ***Ø Se la pelo****,*	… (**I**) **peel it** for her,
	y ***Ø le quito*** *así el piquito,*	and (**I**) **pull out** the tip,'

(CCCS 01: 1494-1496)

(18)

Wood: *(H) and* ***he stood*** *opposite me,*
and ***Ø looked*** *at me,*

(SBCSAE 55: 161-162)

(19)

Dora:	*Pero* ***ella lloraba*** *mucho.*	'But **she would cry** a lot.'
Ángela:	*a=h.*	'Oh.'
Dora:	***Ø Decía*** *!ay mis nenas,*	'**(She) would say** oh my girls.'

(CCCS 30: 595-597)

(20)

Darlene: *... (TSK) anyway,*
he almost died.
(H)= Too.
He almost died *twice.*

(SBCSAE 52: 1066-1068)

[15] Thus excluded were instances of preceding coreferential subjects in subordinate clauses (and in Spanish, also targets unless both the target and preceding occurred as part of the same subordinate clause). An exception to this were temporal clauses, as in the example below (cf. Thompson 1987: 445).

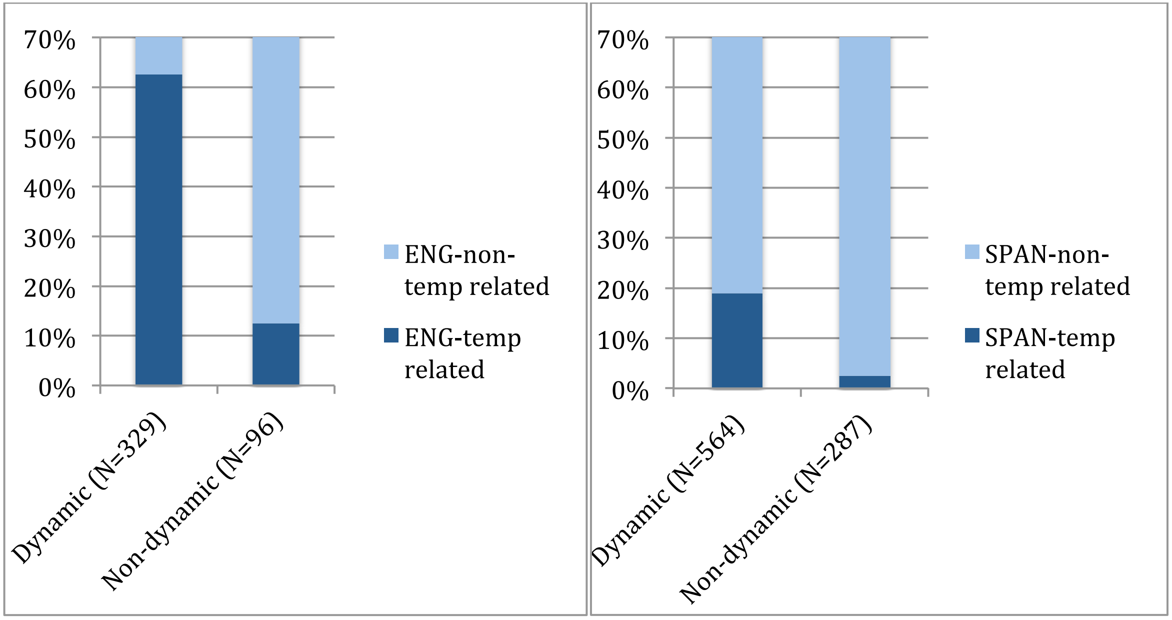
Javier	*.. Cuando Ø vio este llavero,*	'.. When (he) saw this keyring,'
Santi	*... Hm.*	'… Hm.'
Javier	*.. Ø me dijo,*	'.. (he) said to me,'

(CCCS 05: 1069-1071)

There is a substantially higher rate of non-expression in temporally related than in non-temporally related clauses in both languages. For English, subjects are left unexpressed over twice as often when there is a temporal relationship than when there isn't (78% (171/218) compared with 35% (74/210)), and, for Spanish, nearly 1.5 times as often (91% (104/114) vs. 68% (509/744)). As we can infer from these rate differences, the role of temporal sequencing in variable subject expression is more important in English than in Spanish, as was also the case with linking (accessibility). In English, Ø is highly disfavored in the absence of a temporal relationship with the preceding coreferential-subject main clause (Travis and Lindstrom 2016: 114).

It is this temporal relationship effect that accounts, at least in part, for the favoring of unexpressed subjects by dynamic verbs. In both languages a far greater proportion of dynamic than non-dynamic (stative and cognition in Table 1) verbs occur in temporally related clauses, as seen in Figure 10. This association between lexical aspect and temporal relationship is consonant with the distribution of lexical and grammatical aspect (dynamic vs. stative verbs, perfective vs. imperfective aspect) in "foregrounded" vs. "backgrounded" clauses that has been proposed to apply cross-linguistically (Hopper 1979: 215-216; cf. Givón 2001: 339).

Figure 10 Distribution of dynamic vs. non-dynamic verbs according to temporal relationship with preceding main clause with a coreferential subject:
English (left panel) and Spanish (right panel)



8.2 Language-specific lexically particular constructions

We find particular constructions for subject expression in both languages, though they are different in each: English [VERB$_i$ *and* Ø VERB$_i$] and Spanish [*yo* 'I' + Cognition VERB$_{1sg}$].

In English, a particular construction strongly favoring unexpressed subjects is one in which coreferential clauses are conjoined with *and* [VERB$_i$ *and* Ø VERB$_i$], illustrated in (21) (see Section 3, on "VP-coordination").

(21)

Beth: *because Anna* ***takes ahold*** *of something like that* ***and Ø doesn't let go****.*
(SBCSAE 31: 1448)

While a range of verbs occur in this construction, two frequent manifestations are with a verb of motion as the first verb, [Motion Verb$_i$ *and* Ø Verb$_i$], and with a verb of speech as the second verb, [Verb$_i$ *and* Ø Verb-of-speech$_i$] (cf., Brinton 1990: 124-125; Hopper 2002: 152; Quirk et al. 1985: 978; Torres Cacoullos and Travis 2014: 31).

The [Motion Verb$_i$ *and* Ø Verb$_i$] construction occurs 59 times in the data, and has a rate of non-expression of over twice the overall average, at 85% (50/59). The most frequent first verb is *go* (N=28), followed by *come* (N=16), with other motion verbs such as *walk* also included. An example with *go* is seen in (22), where the two conjuncts occur in a single IU, and with *come* in (6b) above, where they occur across IUs in a prosodically linked context (with continuing intonation).

(22)

Danny: *... So she* ***goes and Ø gets****= .. the mother.*
(SBCSAE 30: 432)

The [Verb$_i$ *and* Ø Verb-of-speech$_i$] construction occurs 51 times in the data, and has a rate of non-expression of 88% (45/51). The most frequent speech verbs are *say* (N=28) and *tell* (N=18). This construction is illustrated in (23) across prosodically linked IUs, and in (4) above, within a single IU.

(23)

Lajuan: *Like* ***he called me last week,***
and Ø said *he wanted to be with me.*
(SBCSAE 44: 1068-1069)

These instances of *and* coordination merit the status of particular constructions by virtue of the relatively high rates of unexpressed subjects and their high frequency. Together they account for nearly three quarters (48/68) of cases of conjuncts in the same IU, and approximately one third (101/328) of all structurally linked (prosodic and/or syntactic) cases.[16] A feature of these two particular constructions is that the vast majority (88/101) of tokens occur in temporally related contexts, as compared with just 39% (129/326) of the coreferential tokens outside these particular constructions. Same-IU occurrences can be considered to be the most tightly linked, expressing a single event (Schiffrin 1981:53-55; Torres Cacoullos and Travis 2014: 27-28). Nevertheless, with these constructions, as with coordination more generally, we find gradience from more to less fixed configurations (as can be seen by comparing examples (21), (22), and (23)). It is usage-based constructions on the more tightly-linked end of the continuum that may have fed the intuited notion of "VP coordination".

For Spanish, we turn to cognition verbs, which strongly disfavor unexpressed subjects. This effect only applies to 1sg, since cognition verbs are rare with 3sg subjects: they represent one fifth of the 1sg data (296/1,378), but under 3% of the 3sg data (39/1,403); and close to 90% (296/335) of the cognition verbs occur with 1sg subjects. Spanish cognition

[16] The total number of [Motion Verb$_i$ *and* Ø Verb$_i$] and [Verb$_i$ *and* Ø Verb-of-speech$_i$] is 101; of these, nine are instances of both, as in (5b).

verbs constitute a class: the disfavoring of Ø—conversely, favoring of subject pronoun *yo* 'I'—generally holds across a number of cognition verbs and particularly at the beginning of a speaker turn (Travis and Torres Cacoullos 2012: 735). There is evidence, then, for a general [*yo* 'I' + Cognition VERB$_{1sg}$] construction. This is centered on two lexically particular constructions, which alone account for over one half of all 1sg cognition verb occurrences: *yo no sé* 'I don't know' (N=102) and *yo creo* 'I believe/think' (N=79). Both qualify as particular constructions having not only a rate of unexpressed subjects substantially below the overall average of 59% unexpressed—*yo no sé* 'I don't know' at 34% (35/102), and *yo creo* 'I believe/think' at 44% (35/79)—but also distinct variation patterns (Travis and Torres Cacoullos 2012: 738-742).

In English, as in Spanish, we observe the same association of cognition verbs with 1sg subjects (they account for approximately one quarter of the 1sg tokens (103/388) vs. 9% of the 3sg tokens (42/490)). Unlike Spanish, English cognition verbs do not behave as a class with respect to subject realization (Travis and Torres Cacoullos 2014: 380-381). We find, then, idiosyncratic behavior of particular constructions, favoring (e.g., English [MOTION VERB$_i$ *and* Ø VERB$_i$]) or disfavoring (e.g., Spanish [*yo* 'I'+ Cognition VERB$_{1sg}$], unexpressed subjects.

Thus, lexically-particular constructions for subject expression are language-specific and a locus of difference between English and Spanish. Spanish has a class of cognition verbs, and a language-particular [*yo* 'I'+ Cognition VERB$_{1sg}$] construction, while English displays particular [VERB$_i$ *and* Ø VERB$_i$] constructions. On the other hand, the differential behavior of the classes of dynamic vs. stative verbs has been shown to be due to temporal relationship, and a candidate for a universal constraint on subject expression cutting across cross-linguistic types.

9 Person

Differential behavior for persons has been described as a feature of "partial null subject languages" (Roberts and Holmberg 2010: 11) such as Finnish.[17] Nevertheless, person effects have been reported in quantitative studies across a range of languages (e.g. for Arabic (Owens et al. 2013: 268; Parkinson 1987: 356), Bislama (Meyerhoff 2009: 311), Cantonese and Russian (Nagy et al. 2011: 141-142), and Auslan (McKee et al. 2011: 388)). Results are difficult to interpret, because human and inanimate subjects are not always distinguished. Yet among those that distinguish human subjects, a favoring of non-expression with 3sg over 1sg subjects has been reported for Spanish (e.g., Orozco 2015: 27; Silva-Corvalán 2001: 166), Brazilian Portuguese (Silveira 2011: 48), and Mandarin (Jia and Bayley 2002: 110).

The same applies to English, and thus here again we find a similarity between the two languages: 3sg favors unexpressed subjects more than does 1sg both in Spanish, with rates of 68% (967/1,413) vs. 50% (692/1,389) respectively, and in English, with rates of approximately 5% (180/3,500) vs. 2% (149/6,600).

[17] Though in conversational data, rates of pronominal expression for 1sg and 3sg human subjects are similar (89%, 1,591/1,794 and 87%, 1,173/1,353, respectively) (Helasvuo p.c., cf., Helasvuo, 2014).

Grammatical person differences in rates of pronominal vs. unexpressed subjects are in part attributable to the existence of 3sg full NPs, the favored 3sg form to introduce a new referent or to return to a prior referent that hasn't been mentioned for some time (cf., Dumont 2016: 84). The impact within the variable context for subject expression is that 3sg subjects—pronominal and unexpressed—tend to cluster together. Dahl similarly finds "clustering" for third person in Swedish conversation and remarks that "once you have started talking about a third person referent, the chance that you will continue doing so also in the following clause is much higher than in the case of egophoric [1st and 2nd person] referents, other things being equal" (2000: 65).

The availability of lexical subjects for 3rd person contributes to distinct contextual distributions of 3sg pronominal and unexpressed subjects. These tend to occur more than 1sg subjects in environments that are propitious to unexpressed subjects. Concomitant with the "clustering" referred to above, in terms of linking (accessibility), 3sg subjects occur more often in structurally linked coreferential contexts than do 1sg subjects, approximately one half of the time for 3sg compared with just one third of the time for 1sg (in Spanish, 711/1,391 vs. 347/1,355). In addition, for semantic class of verbs, 3sg subjects occur rarely with cognition verbs which favor pronominal subjects in Spanish (and are overwhelmingly in 1sg, as noted, Section 8.2).

In sum, we hypothesize, on the one hand, that grammatical person differences in subject expression at least in part reflect different distributions according to contexts impinging on subject expression. Once lexical forms are set aside, 3sg subjects generally occur in contexts favorable to non-expression more than 1sg subjects do. Cross-linguistic differences, on the other hand, may be identified in language-particular constructions, such as the class of 1sg cognition-verb constructions in Spanish, which disfavor unexpressed subjects.

10 Conclusion: Using inherent variability to locate universals

Variationist Typology seeks to uncover forces giving rise to grammatical similarities or differences, establishing these through the structure of variation as revealed by systematic quantitative analysis of speech data. Grammatical (dis)similarity is detected not by the presence or absence of a feature, nor by its overall rate of use. The loci of cross-language comparisons are instead both the probabilistic constraints and the variable context within which they are operative. The linguistic conditioning of variation provides operationalizations of surmised communicative or interactional functions and cognitive or processing aspects widely appealed to in language typology.

Rates of use are not a reliable comparison measure. Despite the conspicuous rarity of unexpressed subjects in English compared with Spanish, there is structured variability within this non-null subject language, which, contrary to cherished belief, displays striking parallels with variation patterns in the null-subject language.

First, VP coordination is neither a discrete nor a distinguishing feature of English. Here, we have applied the theoretical construct to actual patterns in speech, operationalizing it as the coordination of coreferential-subject clauses with one particular conjunction, *and*. In the spontaneous speech data studied here, of all conceivable incarnations of VP coordination, *and*-coordination is by far both the most frequent and the most propitious for unexpressed subjects. Furthermore, prosody works with syntax, such that unexpressed subjects under syntactic linking (via *and*) are further favored when the conjuncts are also prosodically linked (when both verbs are in the same Intonation Unit or separated by at most a continuing intonation contour). This reveals a continuum of coordination that contradicts the notion of VP coordination as a discrete category: higher rates of unexpressed subjects with both prosodic and syntactic linking to the preceding coreferential-subject clause, lower rates with either one or the other kind of structural linking between coreferential-subject clauses, and even lower rates in coreferential contexts lacking structural linking. Lowest rates overall are found in non-coreferential contexts. We observed here the same graded effect for Spanish, demonstrating, then, a similarity across the languages in what has been declared a feature specific to English.

Second, this shared probabilistic constraint of linking with the preceding subject demonstrates that the effect of accessibility, long put forward as the most important determinant of subject expression across null-subject languages, also holds for English (and in fact is a stronger constraint than in Spanish). The discovery of a role specifically for linking suggests that accessibility of the subject referent may be reconceived as a composite of both semantic and structural features, to encompass not only coreferentiality with the preceding clause subject but also structural—prosodic and syntactic—linking.

Third, in addition to accessibility, there are two other candidate cross-language constraints on subject expression, which based on the present results we hypothesize will apply independently of classifications of language types. One is coreferential subject priming, whereby speakers favor the form of the preceding coreferential subject that they produced. The other is lexical aspect, such that dynamic predicates favor unexpressed subjects more than stative verbs do; this effect reflects the contribution of a temporal relationship to subject expression, such that unexpressed subjects are favored with situations which are sequential or simultaneous with respect to that of an adjacent main clause with a coreferential subject.

Where, then, does the difference between English and Spanish lie? We have seen that the languages are distinguished by their envelopes of variation. In English, besides coreferential-subject verbs conjoined with a coordinating conjunction, unexpressed subjects are limited to prosodic-initial position in declarative main clauses, a restriction that is absent in Spanish. A second site of cross-language difference lies in lexically-particular constructions, [Motion Verb$_{i}$ *and* Ø Verb$_{i}$] and [Verb$_{i}$ *and* Ø Verb-of-speech$_{i}$] in English vs. the [*yo* + Cognition Verb$_{1sg}$] construction in Spanish. This pair of results suggests the hypothesis that a primary locus of cross-language differences is language-specific variable contexts and lexically-particular constructions.

The results demonstrate that cross-linguistic comparisons will profitably examine not only the probabilistic constraints, but also the envelope of variation within which those constraints operate. These findings also suggest a role for prosody in shaping cross-linguistic morphosyntactic patterns, consideration of which necessitates prosodically-transcribed speech data as the basis for comparative analysis. And they highlight the importance of language-specific, lexically particular constructions, which are best made evident to the researcher in spontaneous speech corpora that provide samples of language as it is actually used by speakers.

The proposals put forward here for typological comparison fit renewed scholarly interest in "quantitative crosslinguistic investigations of discourse" (Haig and Schnell 2016: 615), with a focus on probabilistic constraints, rather than "hard" grammatical rules (e.g., Bresnan, et al. 2001). These lines of work can be advanced, we hope to have shown, by drawing on quantitative analysis of linguistic variation in comparable speech corpora to provide “higher descriptive precision” of language types (Bickel 2015: 921). Variationist Typology thus puts flesh and bones on the notion that language-particular analysis and cross-language comparisons of particular languages are both “part of a larger, coherent endeavor, that of documenting and understanding linguistic diversity” (Haspelmath 2010: 682). In sum, the understanding of language universals can be advanced by examining speaker choices in their spontaneous interactions thus locating *cross*-linguistic differences and similarities in quantitative patterns of *intra*-linguistic variation.

Appendix 1: Mixed effects modeling

Generalized linear mixed effects models (GLMM) were built with a logistic link function using glmer() from the *R* (R Core Team 2015) package (Bates et al. 2015), predicting the non-expression of subject pronouns given Linking, Priming, Verb Class, and Tense, as for the variable rule analyses (VRAs) presented in Table 1. In the GLMM Speaker and Verb were included as random effects to check to what extent the predictors' effects are stable across individual speaker and verb (Baayen 2008). The model summaries are presented below in Table 2 for English and Table 3 for Spanish. Positive coefficients in Table 2 and Table 3 (and probability values closer to 1 in Table 1) indicate an increased likelihood of non-expression and negative coefficients (and probability values closer to 0 in Table 1) indicate an increased likelihood of pronominal expression. The GLMM results are best considered together with the VRA results in Table 1. Table 1 highlights direction and magnitude of effect through probabilities and rates for each predictor level (none of which is singled out as a reference level, unlike the default dummy coding in R applied here for the GLMM). It also reports the data distributions by predictor level, seen in the number of tokens and the percentage of data each level represents.

For a mixed logistic regression to work, sufficient observations are needed (on low token counts inflating individual differences, see Guy, 1980: 15-26). What counts as sufficient may depend on the linguistic variable, in particular the complexity of the linguistic conditioning. In natural speech, unlike experimental, data, distributions are not controlled, and thus a mixed GLM is restricted in the data points it can take into account. Here, just excluding cases where there were fewer than five data points for either verb or speaker resulted in a total of 2,113 data points for Spanish (from a total of 2,802) and 418 for English (from a total of 878).[18]

Overall, the GLMM results are consistent with those of the VRA: in both, Linking and Previous realization are found to have a significant effect, and in neither is Tense significant. A difference is that, of the effects for semantic class in Spanish, it is only that of Dynamic verbs that is significant. While this is the case with the inclusion of the random effect for verb and the reference level set to Stative verbs, significance for Spanish Cognition verbs was achieved in an identical GLMM with Dynamic verbs set as the reference level. These results can only be interpreted by supplementing regression analysis with detailed quantitative views of the data, allowing for constructions and classes of items to be identified. From a linguistic perspective, what is important is that categories are anchored in frequent lexical items, as demonstrated in Section 8.2 (cf., Bybee 2010: Chapter 5).[19]

[18] A higher cut off, e.g. of 30 or more, may provide more meaningful results for individual verb and speaker effects, but this would leave fewer than half the tokens for Spanish (1,236) and none for English. Of the total of 88 speakers in the English sample, 53 speakers produce only one or two unexpressed tokens, 10 speakers have 10 or more, and none have over 25.

[19] While the verbs *saber* 'know' and *creer* 'think' account for a substantial proportion of the cognition verbs, the favoring of subject pronoun *yo* 'I' holds across a number of cognition verbs (Travis & Torres Cacoullos 2012: 735).

Table 2 Generalized linear (mixed) model predicting an unexpressed subject: English (speaker / verb 5+ tokens)

	B	σ	*Z* value	*p* value
(Intercept)	-2.07	0.61	-3.37	0.00
Linking – maximally linked	2.08	0.36	5.78	0.00
Linking – non-coreferential	-0.91	0.40	-2.28	0.02
Priming – previous “other”	0.17	0.34	0.51	0.61
Priming – previous unexpressed	0.84	0.37	2.26	0.02
Verb Class – Cognition	0.95	0.75	1.27	0.20
Verb Class – Dynamic	0.43	0.53	0.80	0.42
Tense – Past	0.26	0.38	0.70	0.49

Overall Ø 37% (153/418); for 34 verb types, Variance = 0.89 (SD = .94) and for 32 Speakers, Variance = 0.00 (SD = 0.00). The zero variance for the speaker random effect, included to maintain parallels with the Spanish model, is because two pronouns were extracted for each unexpressed subject by speaker (Section 5.1).

Table 3 Generalized linear (mixed) model predicting an unexpressed subject: Spanish (speaker / verb 5+ tokens)

	B	σ	*Z* value	*p* value
(Intercept)	0.25	0.20	1.24	0.21
Linking – maximally linked	0.72	0.17	4.19	0.00
Linking – non-coreferential	-0.42	0.11	-3.73	0.00
Priming – previous pronoun	-0.36	0.12	-2.95	0.00
Priming – previous unexpressed	0.68	0.12	5.84	0.00
Verb Class – Cognition	-0.38	0.25	-1.54	0.12
Verb Class – Dynamic	0.56	0.18	3.16	0.00
Tense – Imperfect	-0.12	0.16	-0.74	0.46
Tense – Preterit	0.14	0.13	1.08	0.28

Overall Ø 56% (1,180/2,113); for 81 verb types, Variance = 0.09 (SD = .31) and for 32 Speakers, Variance = 0.18 (SD = 0.43).

Appendix 2: Transcription Conventions (Du Bois et al. 1993)

.	final intonation contour	=	lengthening
,	continuing intonation contour	[ ]	speech overlap
?	appeal intonation contour	!	booster: emphatic speech
--	truncated intonation contour	%	glottal stop
-	truncated word	(H)	in-breath
..	short pause (about 0.5 seconds)	(TSK)	click
...	medium pause (> 0.7 seconds)	(THROAT)	throat clearing

*Acknowledgments

We thank Harald Baayen and the anonymous *Linguistics* reviewers, as well as colleagues who responded to presentations at the International Conference on Language Variation in Europe (ICLaVE) 8, Universität Leipzig, May 2015; New Ways of Analyzing Variation (NWAV) 44, University of Toronto, October 2015; and Australian Linguistic Society (ALS), Western Sydney University, December 2015. We also thank Danielle Barth and Simón González for running the mixed effects models. Both authors contributed equally to this work, made possible by funding from the National Science Foundation (1019112/1019122) and the Australian Research Council Centre of Excellence for the Dynamics of Language.

References

Akmajian, Adrian & Frank Heny. 1980. *An introduction to the principles of transformational syntax*, Cambridge, MA: The MIT Press.

Amaral, Patricia Matos & Scott A. Schwenter. 2005. Contrast and the (non-) occurrence of subject pronouns. In David Eddington (ed.), *Selected proceedings of the 7th Hispanic Linguistics Symposium*, 116-127. Somerville, MA: Cascadilla Press.

Ariel, Mira. 1994. Interpreting anaphoric expressions: A cognitive versus a pragmatic approach. *Journal of Linguistics* 30(1). 3-42.

Baayen, R. Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Barbosa, Pilar, Maria Eugenia Lamoglia Duarte & Mary Aizawa Kato. 2005. Null subjects in European and Brazilian Portuguese. *Journal of Portuguese Linguistics* 4(11-52).

Bates, Douglas, Martin Maechler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1-48, doi:10.18637/jss.v18067.i18601.

Bentivoglio, Paola. 1987. *Los sujetos pronominales de primera persona en el habla de Caracas*, Caracas: Universidad Central de Venezuela.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Edward Finegan & Susan Conrad. 1999. *The Longman grammar of spoken and written English*, London: Longman.

Bickel, Balthasar. 2015. Distributional typology: Statistical inquiries into the dynamics of linguistic diversity. In Bernd Heine & Heiko Narrog (eds.), *The Oxford Handbook of Linguistic Analysis* (2nd ed.), 901-923. Oxford: Oxford University Press.

Bock, J. Kathryn & Zenzi M. Griffin. 2000. The persistence of structural priming: Transient activation or implicit learning. *Journal of Experimental Psychology: General* 129(2). 177-192.

Bresnan, Joan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston & Wolfgang Sternefeld (eds.), *Roots: Linguistics in Search of Its Evidential Base*, 75–96. Berlin: Mouton de Gruyter.

Bresnan, Joan, Shipra Dingare & Christopher D. Manning. (2001). Soft constraints mirror hard constraints: Voice and person in English and Lummi. In Miriam Butt and Tracy Holloway King (Eds.) *Proceedings of the LFG01 Conference*, 13-31. Stanford: CSLI Publications.

Brinton, Laurel. 1990. The development of discourse markers in English. In Jacek Fisiak (ed.), *Historical linguistics and philology*, 45-71. Berlin / New York: Mouton de Gruyter.

Bybee, Joan. 2009. Language universals and usage-based theory. *Language Universals*. 17-39.

Cameron, Richard. 1994. Switch reference, verb class and priming in a variable syntax. *Papers from the Regional Meeting of the Chicago Linguistic Society: Parasession on variation in linguistic theory* 30(2). 27-45.

Cameron, Richard. 1995. The scope and limits of switch reference as a constraint on pronominal subject expression. *Hispanic Linguistics* 6-7. 1-27.

Cameron, Richard & Nydia Flores-Ferrán. 2003. Perseveration of subject expression across regional dialects of Spanish. *Spanish in Context* 1(1). 41-65.

Cedergren, Henrietta & David Sankoff. 1974. Variable rules: Performance as a statistical reflection of competence. *Language* 50. 333-355.

Chafe, Wallace. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Charles N. Li (ed.), *Subject and topic*, 25-55. New York: Academic Press.

Chafe, Wallace. 1988. Linking intonation units in Spoken English. In John Haiman & Sandra A. Thompson (eds.), *Clause combining in grammar and discourse*, 1-27. Amsterdam / Philadelphia: John Benjamins.

Chafe, Wallace. 1994. *Discourse, consciousness and time: The flow and displacement of conscious experience in speaking and writing*, Chicago: University of Chicago Press.

Chambers, J.K. 2004. Dynamic typology and vernacular Universals. In: Bernd Kortmann, (ed.), *Dialectology meets typology: Dialect grammar from a cross-linguistic perspective*, 127-146. Berlin/New York: Mouton de Gruyter.

Chociej, Joanna. 2011. Polish null subjects: English influence on heritage Polish in Toronto, Department of Linguistics, University of Toronto., Ms.

Claes, Jeroen. 2011. ¿Constituyen las Antillas y el Caribe continental una sola zona dialectal? Datos de la variable expresión del sujeto pronominal en San Juan de Puerto Rico y Barranquilla, Colombia. *Spanish in Context* 8(2):191-212.

Comajoan, Llorenc. 2006. Continuity and episodic structure in Spanish subject reference. In J. Clancy Clements & Jiyoung Yoon (eds.), *Functional approaches to Spanish syntax: Lexical semantics, discourse and transitivity*, 53-79. New York: Palgrave MacMillan.

Croft, William. 2001. *Radical construction grammar: Syntactic theory in typological perspective*, Oxford: Oxford University Press.

Dahl, Östen. 2000. Egophoricity in discourse and syntax. *Functions of Language* 7(1). 37-77.

Dixon, R. M. W. 2005. *A semantic approach to English grammar* (2nd ed.), Oxford: Oxford University Press.

Dryer, Matthew S. 2013. Expression of Pronominal Subjects. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online* Leipzig: Max Planck Institute for Evolutionary Anthropology (Available online at http://wals.info/chapter/101, Accessed on 2015-12-18.)

Du Bois, John W., Wallace L. Chafe, Charles Myer, Sandra A. Thompson, Robert Englebretson & Nii Martey. 2000-2005. Santa Barbara corpus of spoken American English, Parts 1-4. Philadelphia, Linguistic Data Consortium.

Du Bois, John W., Stephan Schuetze-Coburn, Susanna Cumming & Danae Paolino. 1993. Outline of discourse transcription. In Jane Edwards & Martin Lampert (eds.), *Talking data: Transcription and coding in discourse*, 45-89. Hillsdale: Lawrence Erlbaum Associates.

Dumont, Jenny. 2016. *Third person references: Forms and functions in two spoken genres of Spanish*, Amsterdam/Philadelphia: John Benjamins Publishing Company.

Enríquez, Emilia V. 1984. *El pronombre personal sujeto en la lengua española hablada en Madrid*, Madrid: Consejo Superior de Investigaciones Científicas, Instituto Miguel de Cervantes.

Erker, Daniel & Gregory Guy. 2012. The role of lexical frequency in syntactic variability: Variation subject personal pronoun expression in Spanish. *Language* 88(3). 526-557.

Evans, Nicholas & Stephen C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. , 32, 429-492.). *Behavioral and Brain Sciences* 32. 429-492.

Ewing, Michael. 2014. Motivations for first and second person subject expression and ellipsis in Javanese conversation. *Journal of Pragmatics* 63. 48-62.

Fernández-Soriano, Olga. 1999. El pronombre personal. Formas y distribuciones. Pronombres átonos y tónicos. In Violeta Demonte & Ignacio Bosque (eds.), *Gramática descriptiva de la lengua española*, 1209–1273. Madrid: Espasa-Calpe.

Givón, T. 1979. *On understanding grammar*, New York: Academic Press.

Givón, T. 1983. Topic continuity in discourse: An introduction. In T. Givón (ed.), *Topic continuity in discourse: A quantitative cross-linguistic study*, 1-41. Amsterdam: John Benjamins.

Givón, T. 2001. *Syntax: An introduction*, vol. 1, 2 vols (2 ed.), Amsterdam: John Benjamins.

Gries, Stefan Th. 2005. Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research* 34(4). 365-399.

Guy, Gregory. 1980. Variation in the group and the individual: The case of final stop deletion. In William Labov (ed), *Language in Time and Space*, 1-36. New York: Academic Press.

Haegeman, Liliane. 1994. *Introduction to Government and Binding Theory*, 2nd ed. Oxford: Wiley-Blackwell.

Haegeman, Liliane. 2013. The syntax of registers: Diary subject omission and the privilege of the root. *Lingua* 130. 88-110.

Haeri, Niloofar. 1989. Overt and non-overt subjects in Persian. *IPrA Papers in Pragmatics* 3:1. 155-166.

Harvie, Dawn. 1998. Null subject in English: Wonder if it exists? *Cahiers Linguistiques d' Ottawa* 16. 15-25.

Haspelmath, Martin. 2004. Coordinating constructions: An overview. *Typological Studies in Language* 58. 3-40.

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86. 663-687.

Helasvuo, Marja-Liisa. 2014. Agreement or crystallization: Patterns of 1st and 2nd person subjects and verbs of cognition in Finnish conversational interaction. *Journal of Pragmatics* 63. 63-78.

Holmberg, Anders. 2010. Null subject parameters. In Theresa Biberauer, Anders Holmberg, Ian Roberts & Michelle Sheehan (eds.), *Parametric variation: Null subjects in minimalist theory*, 88-124. Cambridge / New York: Cambridge University Press.

Hopper, Paul J. 1979. Aspect and foregrounding in discourse. In T. Givón (ed.), *Discourse and syntax*, 213-241. New York: Academic Press.

Hopper, Paul J. 2002. Hendiadys and auxiliation in English. In Joan Bybee & Michael Noonan (eds.), *Complex sentences in grammar and discourse*, 145-174. Amsterdam: John Benjamins.

Huddleston, Rodney & Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*, Cambridge: Cambridge University Press.

Izre'el, Shlomo 2005. Intonation Units and the Structure of Spontaneous Spoken Language: A View from Hebrew. In Cyril Auran, Roxanne Bertrand, Catherine Chanet, Annie Colas, Albert Di Cristo, Cristel Portes, Alain Reynier & Monique Vion (eds.), *Proceedings of the IDP05 International Symposium on Discourse-Prosody Interfaces* CD ROM.

Jia, Li & Robert Bayley. 2002. Null pronoun variation in Mandarin Chinese. *University of Pennsylvania Working Papers in Linguistics* 8(3). 103-116.

Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45(4). 715-762.

Labov, William. 1994. *Principles of linguistic change: Internal factors*, vol. 1, 3 vols, Oxford: Basil Blackwell.

Labov, William. 2005. Quantitative reasoning in linguistics. In Ulrich Ammon, Norbert Dittmar, Klaus J. Mattheier & Peter Trudgill (eds.), *Sociolinguistics/Soziolinguistik: An international handbook of the science of language and society*, vol. 1, 6-22. Berlin: Mouton de Gruyter.

Labov, William & Joshua Waletzky. 1997 [1967]. Narrative analysis: Oral versions of personal experience. *Journal of Narrative and Life History* 7(1/4). 3-38.

Lastra, Yolanda & Pedro Martín Butragueño. 2015. Subject pronoun expression in oral Mexican Spanish. In Ana M. Carvalho, Rafael Orozco & Naomi Lapidus Shin (eds.), *Subject pronoun expression in Spanish: A cross-dialectal perspective*, 39-57. Georgetown: Georgetown University Press.

Lee, Duck-Young & Yoko Yonezawa. 2008. The role of the overt expression of first and second person subject in Japanese. *Journal of Pragmatics* 40(4). 733-767.

Leroux, Martine & Lidia-Gabriela Jarmasz. 2005. A study about nothing: Null subjects as a diagnostic of the convergence between English and French. *University of Pennsylvania Working Papers in Linguistics* 12(2). 1-14.

Levinson, Stephen C. 1987. Pragmatics and the grammar of anaphora: A partial pragmatic reduction of binding and control phenomena. *Journal of Linguistics* 23(2). 379-434.

Li, Charles & Sandra A. Thompson. 1979. Third-person pronouns and zero-anaphora in Chinese discourse. *Syntax and Semantics* 12. 311-335.

Matthews, P.H. 1981. *Syntax*, Cambridge: Cambridge University Press.

McKee, Rachel, Adam Schembri, David McKee & Trevor Johnston. 2011. Variable “subject” presence in Australian Sign Language and New Zealand Sign Language. *Language Variation and Change* 23. 375-398.

Meyerhoff, Miriam. 2009. Replication, transfer, and calquing: Using variation as a tool in the study of language contact. *Language Variation and Change* 21(3). 297-317.

Miller, Jim. 1995. Does spoken language have sentences? In F. R. Palmer (ed.), *Grammar and meaning: Essays in honour of Sir John Lyons*, 116-135. Cambridge: Cambridge University Press.

Mithun, Marianne. 1988. The grammaticization of coordination. In John Haiman & Sandra A. Thompson (eds.), *Clause combining in grammar and discourse*, 331-359. Amsterdam: John Benjamins.

Nagy, Naomi G., Nina Aghdasi, Derek Denis & Alexandra Motut. 2011. Null subjects in heritage languages: Contact effects in a cross-linguistic context. *University of Pennsylvania Working Papers in Linguistics* 17(2). http://repository.upenn.edu/cgi/viewcontent.cgi?article=1202&context=pwpl.

Napoli, Donna Jo. 1982. Initial material deletion in English. *Glossa* 16. 85-111.

Oh, Sun-Young. 2007. Overt reference to speaker and recipient in Korean. *Discourse Studies* 9(4). 462–492.

Orozco, Rafael. 2015. Pronominal variation in Colombian Costeño Spanish. In Ana M. Carvalho, Rafael Orozco & Naomi Lapidus Shin (eds.), *Subject pronoun expression in Spanish: A cross-dialectal perspective*, 17-37. Georgetown: Georgetown University Press.

Otheguy, Ricardo & Ana Cecilia Zentella. 2012. *Spanish in New York: Language contact, dialect levelling, and structural continuity*, Oxford: Oxford University Press.

Otheguy, Ricardo, Ana Cecilia Zentella & David Livert. 2007. Language and dialect contact in Spanish of New York: Toward the formation of a speech community. *Language* 83(4). 770-802.

Owens, Jonathan, Robin Dodsworth & Mary Kohn. 2013. Subject expression and discourse embeddedness in Emirati Arabic. *Language Variation and Change* 25(2). 255-285.

Paredes Silva, Vera Lucia. 1993. Subject omission and functional compensation: Evidence from written Brazilian Portuguese. *Language Variation and Change* 5(1). 35-49.

Parkinson, Dilworth B. 1987. Constraints on the presence/absence of 'optional' subject pronouns in Egyptian Arabic. *15th Annual Conference on New Ways of Analyzing Variation*. 348-360.

Payne, Thomas E. 1997. *Describing morphosyntax: A guide to field linguists*, Cambridge: Cambridge University Press.

Poplack, Shana & Marjory Meechan. 1998. Introduction: How languages fit together in codemixing. *International Journal of Bilingualism* 2(2). 127-138.

Poplack, Shana & Sali Tagliamonte. 1999. The grammaticization of *going to* in (African American) English. *Language Variation and Change* 11(3). 315-342.

Poplack, Shana, Lauren Zentz & Nathalie Dion. 2012. What counts as (contact-induced) change. *Bilingualism: Language and Cognition* 15(2). 247-254.

Posio, Pekka. 2013. The expression of first-person-singular subjects in spoken Peninsular Spanish and European Portuguese: Semantic roles and formulaic sequences. *Folia Linguistica* 47(1). 253-291.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*, London: Longman.

R Core Team. 2015. *R: A language and environment for statistical computing*, Vienna, Austria: R Foundation for Statistical Computing, http://www.R-project.org.

Rizzi, Luigi. 1982. *Issues in Italian syntax*, Dordrecht: Foris publications.

Roberts, Ian & Anders Holmberg. 2010. Introduction: Parameters in minimalist theory. In Theresa Biberauer, Anders Holmberg, Ian Roberts & Michelle Sheehan (eds.), *Parametric variation: Null subjects in minimalist theory*, 1-57. Cambridge: Cambridge University Press.

Rögnvaldsson, Eiríkur. 1982. We need (some kind of a) rule of conjunction reduction. *Linguistic Inquiry* 13. 557-561.

Sankoff, David. 1988. Variable rules. In Ulrich Ammon, Norbert Dittmar & Klaus J. Mattheier (eds.), *Sociolinguistics: An international handbook of the science of language and society*, vol. 2, 984-997. Berlin: Walter de Gruyter.

Sankoff, David, Sali Tagliamonte & Eric Smith. 2012, Goldvarb LION: A variable rule application for Macintosh, University of Toronto, URL http://individual.utoronto.ca/tagliamonte/goldvarb.htm.

Scheibman, Joanne. 2001. Local patterns of subjectivity in person and verb type in American English conversation. In Joan Bybee & Paul J. Hopper (eds.), *Frequency and the emergence of linguistic structure*, 61-89. Amsterdam: John Benjamins.

Schiffrin, Deborah. 1981. Tense variation in narrative. *Language* 57(1). 45-62

Shin, Naomi Lapidus. 2014. Grammatical complexification in Spanish in New York: 3sg pronoun expression and verbal ambiguity. *Language Variation and Change* 26(3). 303-330.

Sigurðsson, Halldór Ármann & Joan Maling. 2010. The empty left edge condition. In Michael Putman (ed.), *Exploring Crash-Proof Grammars* Amsterdam: John Bejnamins.

Silva-Corvalán, Carmen. 1994. *Language contact and change: Spanish in Los Angeles*, Oxford: Clarendon Press.

Silva-Corvalán, Carmen. 2001. *Sociolingüística y pragmática del español* (Georgetown Studies in Spanish Linguistics), Washington, DC: Georgetown University Press.

Silva-Corvalán, Carmen. 2003. Otra mirada a la expresión del sujeto como variable sintáctica. In Francisco Moreno Fernández, Francisco Gimeno Menéndez, José Antonio Samper, María Luz Gutiérrez Araua, María Vaquero & César Hernández (eds.), *Lengua, Variación y contexto: Estudios dedicados a Humberto López Morales*, vol. 2, 849-860. Madrid: Arco Libros.

Silveira, Agripino S. 2011. Subject expression in Brazilian Portuguese: Construction and frequency effects. PhD thesis, Department of Linguistics, University of New Mexico.

Sorace, Antonella. 2004. Native language attrition and developmental instability at the syntax–discourse interface: Data, interpretations and methods. *Bilingualism: Language and Cognition* 7(2). 143-145.

Szmrecsanyi, Benedikt. 2005. Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* 1(1). 113-149.

Thompson, Sandra A. 1987. 'Subordination' and narrative event structure. In Russell S. Tomlin (ed.), *Coherence and grounding in discourse*, 435-454. Amsterdam / Philadelphia: John Benjamins.

Toribio, Almeida Jacqueline. 2000. Setting parametric limits on dialectal variation in spanish. *Lingua: International Review of General Linguistics* 110. 315-341.

Torres Cacoullos, Rena & Catherine E. Travis. 2014. Prosody, priming and particular constructions: The patterning of English first-person singular subject expression in conversation. *Journal of Pragmatics* 63. 19-34.

Torres Cacoullos, Rena & Catherine E. Travis. 2018. *Bilingualism in the community*. Cambridge: Cambridge University Press.

Travis, Catherine E. 2005. *Discourse markers in Colombian Spanish: A study in polysemy* (Cognitive Linguistics Research), Berlin / New York: Mouton de Gruyter.

Travis, Catherine E. 2007. Genre effects on subject expression in Spanish: Priming in narrative and conversation. *Language Variation and Change* 19(2). 101-135.

Travis, Catherine E. & Amy M. Lindstrom. 2016. Different registers, different grammars? Subject expression in English conversation and narrative. *Language Variation and Change* 28 (1): 103-128

Travis, Catherine E. & Rena Torres Cacoullos. 2012. What do subject pronouns do in discourse? Cognitive, mechanical and constructional factors in variation. *Cognitive Linguistics* 23(4). 711-748.

Travis, Catherine E. & Rena Torres Cacoullos. 2014. Stress on *I*: Debunking unitary contrast accounts *Studies in Language* 38(2). 360-392.

Travis, Catherine E., Rena Torres Cacoullos & Evan Kidd. 2015. Cross-language priming: A view from bilingual speech *Bilingualism: Language and Cognition (Special issue edited by Gerrit Jan Kootstra and Pieter Muysken)* doi:10.1017/S1366728915000127.

Weinreich, Uriel, William Labov & Marvin I. Herzog. 1968. Empirical foundations for a theory of language change. In Winfred P. Lehmann & Yakov Malkiel (eds.), *Directions for historical linguistics: A symposium*, 95-188. Austin, TX: University of Texas Press.

Weir, Andrew. 2012. Left-edge deletion in English and subject omission in diaries. *English Language and Linguistics* 16(1). 105-129.

Wolk, Christoph, Joan Bresnan, Anette Rosenbach & Benedikt Szmrecsanyi. 2013. Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica* 30(3). 382–419. doi:10.1075/dia.30.3.04wol.