Referential Null Subjects (RNS) in colloquial spoken Persian: Does speaker familiarity have an impact?[1]

Geoffrey Haig Širin Adibifar

University of Bamberg

Pre-publication draft, March 2019, to appear in:

Korangy, Alireza and Behrooz Mahmoodi-Bakhtiari (eds.) *Essays on the typology of Iranian languages*. Berlin: Mouton de Gruyter.

1 Referential Null Subjects (RNS): Background

It is well known that languages differ considerably in the extent to which a clause requires an overt subject NP. Some languages, like Persian, tolerate clauses without overt subject constituents in a very wide range of contexts, while others (e.g. English) only permit referential null subjects (RNS) under highly constrained conditions. In the literature, two approaches to these cross-linguistic differences can be discerned: a parametric approach, and a discourse, or usage-based approach. The parametric approach goes back to Perlmutter (1971), who introduced a 'pro-drop parameter', according to which a language requires, or does not require, overt expression of referential subjects. The Pro-drop parameter was exclusively concerned with zero subjects, but the original either/or Pro-drop Parameter has since given way to more refined typologies, involving four distinct types (e.g. Holmberg 2009), and has been extended to include Referential Null Objects under the label "radical pro-drop" or "discourse pro-drop" (e.g. Neeleman & Szendröi 2008). Within parametric approaches, the presence of RNS is often linked to the presence of rich agreement morphology, which apparently licenses RNS, though what exactly constitutes rich agreement continues to be a matter of controversy (see Camacho 2013 for recent discussion of null subjects and rich agreement).

While research in the parametric tradition continues to perceive RNS as a parameter of individual grammars, a second line of research is usage, or discourse based. On this view, RNS is a locus of gradual variation, thus not entirely determined by 'the grammar' of a language, but also

[1] We would like to express our gratitude to the 29 Persian speakers who contributed data to this study, to the Language Archive of Cologne (LAC) for hosting the corpus, to Nils Schiborr for data handling, and the Multi-CAST team for collaboration on annotation and corpus design. We would also like to thank audiences in Cologne and Bamberg for critical feedback on earlier presentations of this research. None of these people bear any responsibility for the remaining shortcomings.

dependent on contextual and interactional factors. Methodologically, this approach adopts empirical, quantitative methodologies, drawing on the analysis of language usage ('performance') rather than on intuitions regarding grammaticality. Within language typology, this line of research is associated with Bickel and associates' work on Referental Density (RD) (Bickel 2003, Stoll and Bickel 2009). RD is an empirical measure of the overall density of overt argument expressions in actual discourse, and is not restricted to subjects and objects. Thus RD is not conceived as a categorical feature of the grammar of "a language", but as a graded value, characterizing a specific stretch of discourse. Unlike pro-drop and its descendants, RD does not yield an either/or parameter setting for a particular language, but exhibits a certain degree of language-internal variation, depending on text type and other factors to be taken up below, and is a tool of corpus-based, rather than grammar-based typology (Haig et al. 2011).

Probably the best-known empirical approach to RNS is within variationist sociolinguistics. Rates of overt vs. zero expression of subjects have been extensively investigated as a linguistic variable, most notably across different varieties of Spanish (see Pešková (2013: 120-121) and Carvalho et al. (2015) for discussion of the relevant literature). Among the main findings of this research is the recognition that dialects of 'the same' language can vary quite considerably, a finding which is of considerable relevance in connection with a large and dialectally diverse language such as Persian. This paper adopts a usage-based perspective on RNS in Persian, drawing on corpus of colloquial spoken Persian and applying quantitative methods to address the issue of which factors are relevant in affecting the rate of RNS in natural discourse. Thus the assumption is that RNS is a variable, rather than categorical feature, and it is the analyst's task to determine the factors which drive the variation.

With the exception of Saeli and Miller (2018),[2] there has been no comparable quantitative research on colloquial spoken Persian. In our study, we focus on the the factor of 'familiarity' between the interlocutors, which has been suggested as relevant in this regard, but we also consider gender of the speaker. Although the current sample is small (see Section 3), our provisional finding is that rates of RNS in Persian are not sensitive to either speaker familiarity, or gender, but in fact emerge as a relatively stable variable across a range of different speakers. These findings echo to some extent the picture from research on better-studied languages, in particular Spanish, which show that with regard to RNS, it is primarily language-internal factors that determine most of the attested variation (see next section).

2 Factors determining rates of RNS: Previous research

The most detailed research on the factors impacting on RNS stems from the variationist sociolinguistic tradition within Hispanic linguistics (see the contributions in Carvalho et al 2015).[3] All

[2] Saeli and Miller (2018) are concerned with the impact of extra-linguistic factors on colloquial spoken Persian, including the issue of pronoun omission, based on elicited responses to a 'favor-asking' task. However, the pronouns concerned in their research are second person forms, rather than the third person forms that dominate in our data. They find an effect of same vs. different gender in speaker diads, but the absolute number of second person subject forms in their data is just 25 (including tokens of the polite pronoun *šomā*, familiar pronoun *to*, and zero, cf. Saeli and Miller 2018: Table 2, p. 180). Nevertheless, this is a promising avenue for future research that complements the current study, both in methodology and the domain of investigation.

[3] Most of these studies use some measure of pronoun omission or retention as the unit for investigating what I have termed RNS. Thus the unit of comparison is not zero subjects as a percentage of all subject NPs (as it is here), but rates of pronominal versus zero subjects. Nevertheless, both measures are ultimately concerned with the same phenomenon, though the resultant figures are not directly comparable.

investigations to date confirm that the primary determinants of subject expression are language internal, with a surprisingly high degree of overlap across different studies with regard to the nature of the relevant factors. The impact of speaker-related factors (e.g. age or gender), on the other hand, has not been consistently demonstrated. Among the linguistic factors, the following are worth mentioning:

Person and number value of the pronoun

This appears to be the highest-ranking factor in determining rates of subject omission. Pešková (2013) notes significantly higher rates of pronoun expression in the first and second person as opposed to the third person, while Carvalho et al. (2015) state that the "broadest generalization" is that singular pronouns are more frequently overt than plural pronouns.

Distance and role of antecedent

As a general finding, subject omission is favoured when the antecedent is subject of the immediately preceding clause, with rates of pronoun retention increasing with increasing distance of the antecedent.

TMA morphology of the verbs

In Spanish, different tense/aspect values are associated with different types of agreement patterns. Carvalho et al. (2015) suggest that pronouns are used more frequently with verb forms with the least ambiguous agreement paradigms.

Lexical semantics of the verb

Peškova (2013), working with elicited data, finds epistemic verbs ('know', 'believe') have higher rates of pronoun retention than perceptive verbs.

The ranking of these factors varies from study to study; nevertheless, there seems to be a broad consensus that person and number of the pronoun is the most predictive factor. Turning to the speaker-related factors, the findings here are less consensual. Several studies find an effect of gender. Alvaraz (2015), based on the Spanish of Santo Domingo, finds a weak preference for pronoun retention among women, as does Orozco (2015) for Colombian Costeño Spanish, though the latter case also shows interaction with age. Peškova (2013), on the other hand, does not mention an effect of speaker gender in her investigation, which, unlike the others discussed in this paragraph, is based on a controlled production experiment. Age is also reported as relevant, but the direction of the correlation is not consistent. Orozco (2015) reports that in Mexico City, younger speakers use fewer overt pronouns, while in Puerto Rican Spanish the opposite trend is found. Genre (arguably an internal factor) is also reported as relevant: argumentation favours overt pronouns, while narration favours pronoun omission (Carvalho et al. (2015).

A further factor that has been discussed in this connection is the degree of familiarity between the interlocutors. Bickel (2011), investigating overall rates of zero argument expression (Referential Density, RD), claims an effect of degree of personal familiarity: where speaker and addressee are personally acquainted, fewer arguments receive overt expression, while lack of personal familiarity leads to higher rates of overt arguments. A related claim is made by Meyerhoff (2011), who discusses rates of subject pronoun deletion in Bislama, the English-based creole of Vanuatu. She compared two versions of the same story, one recounted by a native speaker to his extended family, and one version recounted by the same speaker to the investigator (i.e. an out-group person). Rates of subject pronoun omission were nine percentage points higher with familiar addressees than with the out-group addressee. Meyerhoff (2011: 45) suggests that the higher frequency of subject pronouns used with the out-group addressee may be motivated by the speaker's desire to provide "a non-native speaker with more overt information about who he is referring to in any given sentence".

These findings point to an intuitively plausible impact of speaker familiarity on rates of argument realization: when speakers are addressing persons with whom they are familiar, they can afford to reduce overt informational density, relying on the shared body of cultural knowledge and the addressee's assumed familiarity with the speaker's speech habits to fill in the gaps. When addressing a stranger, however, the speaker cannot assume shared cultural knowledge and familiarity with routinized speech habits, and will accordingly switch to a more explicit style, leading to an overall higher level of overt argument expression. If speaker familiarity is indeed a factor in affecting rates of overt vs. zero subject expression, this would be in line with approaches to linguistic variation which focus on accommodation to the addressee, such as Bell's "Audience Design" (Bell 2006).

3 Research question and data

Persian is a southwest Iranian language of the Indo-European family, and the official language of the Islamic Republic of Iran.[4] It exhibits a mixed word-order typology, with OV order in the clause, but with head-initial ordering elsewhere. With regard to RNS, it has been claimed that "Persian is also a radical pro drop language with frequent use of null arguments in both subject and object positions" (Sato and Karimi 2016: 3). However, we are unaware of any empirically-based approaches to RNS in Persian to date. In this paper, we investigate RNS in a corpus of spontaneous spoken Persian and investigate the role of a number of linguistic and speaker-related factors. The main focus is on the factor of speaker familiarity, as discussed in the preceding section: Do speakers tend to use more RNS when they are personally familiar with their interlocutors?

Although there is no previous research on this specific issue in Persian, we nevertheless considered that Persian could be a potentially interesting laboratory for investigating factors such as interlocutor familiarity, because Persian is characterized by an elaborated range of registers and styles. Speakers are highly sensitive to degrees of formality and to politeness norms, adapting phonology, lexical choices, address forms, and grammar accordingly (Jahangiri 1980, Saeli and Miller 2018). Thus it seemed a reasonable hypothesis that in a language community where speech habits are intimately tied to social status and familiarity, the likelihood of an effect of speech setting on RNS would be high. In order to test this, we compiled a corpus of spontaneous spoken Persian (see next section), under conditions that were controlled for speaker familiarity, and analysed the resulting data quantitatively.

Finally, we note that in Persian, finite verbs obligatorily agree with their subjects via a set of six distinct person and number suffixes on the verb. There is a set of free pronouns, which may be omitted under conditions of pragmatic recoverability, and which are flagged for syntactic function in the same manner as nouns (i.e. with the accusative clitic =*rā*, or via various prepositions). With respect to 'pro-drop', then, these are the relevant pronouns. Verbs do not agree obligatorily with objects, though objects may be indexed on the verb through a set of clitic pronouns. The clitic pronouns are briefly mentioned in connection with certain predicate types in examples (4)-(6) below, but are otherwise

[4] We continue to use the term traditionally used in the western academic tradition "Persian", although the speakers refer to their language as Farsi.

not relevant here (see Rasekh 2014, Mahootian and Gebhardt 2018, and Haig, under review, for discussion of clitic pronouns and agreement).

4 Experiment design and setting

The aim of the study is to test whether speaker familiarity has a significant impact on RNS. In order to test this, we gathered data from 29 native speakers of Persian, with the speakers divided into two groups on the basis of their degree of familiarity with the interviewer (who remained the same throughout).[5] One group included only persons who were either connected to the interviewer through a kinship relationship, or a close personal friendship of at least two years. Interviews with this group took place in a relaxed domestic setting in the region of the interviewer's home town in the Mazanderan region of northern Iran, and in three cases in the speakers' apartments in south Germany. Respondents from the second group had no prior contact to the interviewer. They were recruited among students via their lecturers from the Islamic Azad University in Tehran and Behšahr University in Mazanderan Province. Interviews with these speakers were conducted in seminar rooms of the respective universities, thus heightening the contrast in settings between the two groups.

All interviews took place entirely in Persian. The methodology largely replicates that of Bickel (2003), though with minor modifications: respondents were shown the Pear Story, a six-minute video clip widely used in cross-linguistic investigations of discourse (Chafe 1980), on a laptop computer, and then asked to recount the story to a native-speaker interviewer.[6] The film contains no speech, but the storyline is simple and can be readily grasped by those watching the film. Pear Story retellings have been widely used in cross-linguistic studies of discourse, so that the resulting corpus of Persian is also of considerable utility for future researchers. The entire corpus with annotations is available under a Creative Commons License Agreement,[7] and is thus available for re-analysis or re-interpretation by other scholars.

The sample of respondents was intended to be representative of educated, young adult, native speakers of standard Persian, socialized in an urban environment. Prior to the recordings, all speakers provided basic information regarding age, gender, education, places of socialization, languages of communication (in and outside of the domestic setting), and language of their parents. Prior to the recordings, all speakers received the same set of instructions in Persian, provided by the interviewer, a female educated native speaker of Persian from the same age cohort.

Recordings were transcribed, translated, and syntactically annotated using the GRAID system, which provides a set of decision procedures for identifying zero arguments (Haig & Schnell 2014: 7-8; Haig & Schnell 2016). Transcriptions, translations, and annotations were entered into the software ELAN, which time-aligns annotations with the sound file.[8]

[5] Originally 30 interviews were conducted, but one speaker did not produce a coherent narrative that would have been comparable to the other texts, and that text was excluded. This left two groups with 15 and 14 speakers respectively (see Appendix A for details).

[6] In this respect, our methodology departs from that of Bickel (2003) and Chafe (1980) in that the respondents recounted the story to the same interviewer who showed them the film, rather than to another person. Given the aims of the experiment, it was crucial to keep the identity of the interviewer constant across all groups in order to reduce the impact of factors outside of the main dependent variable, that of speaker familiarity.

[7] See https://lac2.uni-koeln.de/en/multicast/

[8] Developed by Han Sloetjes at the MPI Nijmegen, see https://tla.mpi.nl/tools/tla-tools/elan/.

4.1Issues in coding and analysis

The concept of 'subject' has been variously defined at different times, and in different approaches to syntax. Whether or not all clauses, in all natural languages, should be analysed in such a way that they 'have' (at some level of analysis) a subject, is an open question. But on the assumption that a very significant number of clauses in a very significant number of languages can be analysed in this manner is sufficient justification for maintaining it as a concept of syntactic theory.[9] We thus follow mainstream practice and assume that subjects can be relatively uncontroversially identified for Persian, though we note some problematic cases below.

The basic unit of analysis is the clause unit, consisting of a predicate plus associated arguments. For each clause unit, the subject constituent is identified and coded as either full (or lexical) NP, pronoun, or zero. Example (1)[10] shows a clause unit with an overt lexical subject NP. Example (2) contains a sequence of three clause units, the first with an overt subject NP and the second and third clauses with zero subjects. Example (3) contains a sequence of clauses with zero subjects (clause (3c) also contains a zero object).

(1)	*bad*	*yek*	*pesar-i*	*mi-yā-yad*
	then	one	boy-INDEF	INDIC-come.PRS-3SG

'then a boy comes by' (g1_f_08/06)

(2)	a.	*in*	*pesar-e*	*bā*	*dočarxe*	*āmad*
		this	boy-DEF	with	bike	come.PST.3SG
	b.	Ø	*rad*	*šod*		
		Ø	passing	become.PST.3SG		
	c.	Ø	*raft*			
		Ø	go.PST.3SG			

[9] Within various versions of Generative Grammar, the subject role is generally derived from a particular structural configuration, for example as the Specifier of an IP in a GB approach (Farrell 2005: 176), or in Minimalism as e.g. a NP that is c-commanded by a finite complementizer (Radford 2004: 136), or via checking of nominal features (Farrell 2005: 181). Within LFG and related theories, the subject role is a non-derived category within the layer of structure known as F-Structure. In less formalized, but typologically-inspired approaches to syntax, various 'cluster-concept' notions of subjecthood have been put forward involving structural, semantic, and information-structure related properties. These were pioneered in Keenan (1976); see Comrie (1989: 104-123) and Falk (2006: 1-21) *inter alia* for discussion. Philippine-type and syntactically ergative languages continue to pose certain challenges for a universal definition of subject, but these lie outside the scope of the present paper.

[10] All examples are sourced according to the group (g1 = familiar speakers, g2 = unfamiliar speakers), gender (m/f), and number of the recording. Abbreviations used in the examples are: ACC = accusative; ADD = additive particle; AUX = auxiliary; DEF=definite; INDEF = indefinite; INDIC = indicative; PL = plural; POSS = possessive; PROG = progressive; PRS = present; PST = past; SG = singular.

a. ‘This boy with the bike came along
b. passed by
c. went.’ (g2_f_06/09)

(3)	a.	*bad*	*Ø*	*mive-hā=rā*	*čid*	
		then	Ø	fruit-PL=ACC	pick.PST.3SG	
	b.	*va*	*Ø*	*āmad*	*pāyin*	*va*
		and	Ø	come.PST.3SG	down	and

c.	*Ø*	*Ø*	*rixt*	*tuye*	*sabad*
	Ø	Ø	pour.PST.3SG	into	basket

a. ‘Then (he) picked the fruit
b. and came down and
c. (he) poured (them) into the basket.’ (g1_m_04/2)

Persian has one type of clause which poses certain difficulties for identifying subjects. Semantically, these involve predicates of perception and cognition. Syntactically, they are typically lexicalized combinations of a light verb and some non-verbal element. The NP expressing the Experiencer, if present in the clause, is in the nominative case, but is obligatorily indexed through a possessive clitic attached to the non-verbal element of the complex predicate. The light verb takes the default 3SG person agreement marker. The commonest expression of this type in our corpus is *havās=aš part šodan* ‘attention=3SG separated become’, i.e. ‘to be distracted’. Examples of experiencer predicates are found in (4), (5b), and (6b):

(4)	*Ø*	*češm =aš*	*in*	*sabad-hā=rā*	*gereft*
	Ø	eye=POSS.3SG	this	basket-PL=ACC	take.PST.3SG

‘(He) caught sight of these baskets (lit. his eye took the baskets)’ (g1_f_05/5)

(5)	a.	*yek*	*doxtarxānum-i*	*dāšt*	*bā*	*dočarxe*	*miy-ām-ad*
		one	girl-INDEF	AUX.PST.3SG	with	bicycle	PROG-come.PST-3SG

b.	*ke*	*Ø*	*havās =aš*	*be*	*u*	*part*	*šod*
	so	Ø	attention=POSS.3SG	to	3SG	separated	become.PST.3SG

a. ‘a girl was coming by on a bike
b. so his attention was distracted to her ...’ (g2_m_08/07)

(6)	a.	*kolāh =aš*	*mi-oft-ad*
		hat =POSS.3SG	INDIC-fall.PRS.3SG

b.	*bad*	*in*	*ham*	*havās =aš*	*part*	*mi-šav-ad*
	then	3SG	ADD	attention=POSS.3SG	separated	INDIC- become.PRS.3SG

a. 'His hat falls off,
b. then he gets distracted' (lit. he his.attention becomes separated)' (g1_f_14/13)

The correct analysis of such constructions is a matter of some debate (see e.g. Ghomeshi, forthcoming). We follow Sedighi (2010) and assume that the experiencer constituents of these predicates are subjects, because they exhibit most of the syntactic characteristics of canonical subjects in Persian, and we therefore include them in the overall counts for subjects. However, they are Non-Canonical in the sense that the nature of the agreement morphology they are associated with differs from the agreement morphology associated with canonical subjects in Persian (see Haig (2008: 19-22) for discussion of Non-Canonical Subjects with reference to Iranian languages). Rather than a verbal affix, the agreement morphology is an obligatory clitic, e.g. *=aš* in (5b), which we thus analyse as non-pronominal in this context. What this means is that in (4) and (5b) we count a zero subject, while in (6), we count the pronoun (actually a proximal demonstrative) *in* as a pronominal subject.[11]

Subordinate clauses, including relative clauses, generally involve finite syntax in Persian and are thus not significantly different from independent clauses. We have therefore included them in the data, but followed the procedure of Bickel (2003) in considering only those subject constituents that could be overtly realized, without impairing grammaticality. Where unequivocal decisions could not be reached, the string was marked as 'nc' (not classifiable), and excluded from the counts.

Rate of RNS (or simply 'RNS') was calculated by dividing the number of zero subject constituents in a given text by the overall number of subjects in that text, yielding a figure between zero and one. For example, the speaker g1_m_1 has an RNS value of 0.558, indicating that somewhat more than half of all clauses in his text contained a Referential Null Subject. The mean value for RNS across all speakers was 0.589; see Appendix A below for details.

4.2Variables and hypotheses

The main dependent variable is zero versus overt expression, or more generally, rates of RNS, calculated as the rates of zero subjects against the total number of subjects produced, yielding values between 0.0 (no subjects are zero) and 1.0 (all subjects are zero). Our main aim was to test the effects of speaker familiarity on rates of RNS, but we also considered a number of other predictor variables. These include two speaker-related factors, and two linguistic variables.

[11] We interpret the 3sg pronoun/demonstrative *in* in this example as an overt pronominal expression of the Non-Canonical Subject in the second clause, triggered by the subject change between the first clause (*kolāh=aš* 'his hat' and the implied subject of the second clause (the boy).

4.2.1 Speaker-related variables

Age and gender

Although the available literature yields no obvious hypothesis regarding the effects of these two variables (cf. Section 2 above), we include them as standard variables in variationist research.

4.2.2 Linguistic variables

Number of clause-units (CU's) in each text

Each text is an individual re-telling of the Pear Story film, produced by one speaker. The different speakers actually produced texts of very varied length, measured as the number of CU's (mean 49, SD 21). Some speakers produced an exceedingly brief, almost telegraphic, re-telling, while others were quite elaborated. We assumed a possible effect of length on rates of RNS, based on the following assumption: Given that these narratives contain approximately the same content, all other things being equal, a longer text would offer greater opportunities for zero expression, because zero expression is connected to discourse persistence; a participant to which repeated reference is made over consecutive clauses is more likely to be coded with zero, hence yielding an overall higher rate of RNS. The initial hypothesis with regard to length, then, is that length correlates with higher rates of RNS.

New Referents per Clause Unit (NewRef/CU)

This variable relates to the notion of "Information Pressure" (Du Bois 1987): texts differ in the extent to which they accommodate new information (the introduction of new referents). Some texts recount the continued actions of a small number of protagonists, while others involve repeated introductions of new referents. The latter are characterized by what Du Bois (1987) refers to as "high information pressure", measured in terms of new referents per clause unit. The general assumption is that higher information pressure would correlate with lower rates of RNS, because new referents involve overt expressions, as opposed to zero (see Stoll & Bickel 2009 for counter-examples, and Haig & Schnell (2016) for critical discussion of Information Pressure). We therefore counted for each text the number of new referents introduced, restricting ourselves to individualized entities introduced in the form of a NP, and potentially pronominalizable, yielding an absolute figure of new referents per text (mean 15, SD 4). We then divided that figure by the number of clause units (cf. preceding variable), yielding the rate of new referent introduction per clause unit for each text. The hypothesis is that high information pressure will correlate negatively with rates of RNS.

5 Results

The absolute figures from the 29 transcribed and coded texts are provided in Appendix A. Figures 1 and 2 show the results of the linguistic variables 'Length of text in CU's' (Fig.1), and 'New Referents per CU' (Fig. 2), while Table 1 provides the Pearson Correlation Tests.

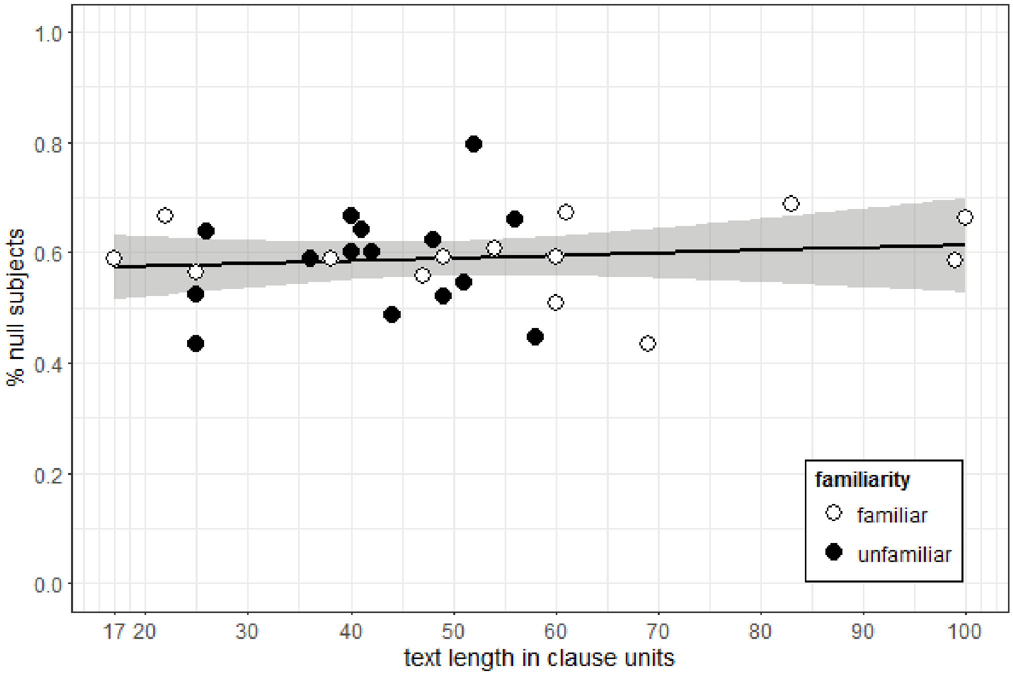

Figure 1: The effect of text length on RNS

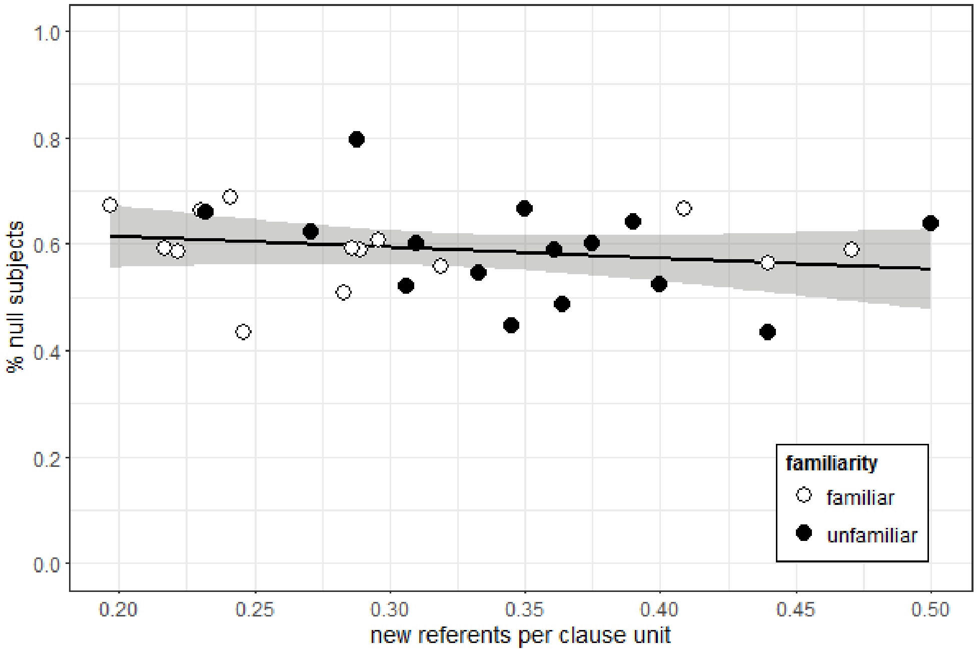

Figure 2: The effect of new referent per CU on RNS

Table 1: Pearson Correlation Tests for Figs. 1 and 2

FACTOR	r	p
RNS | Text length:	0.1212	0.531
RNS | newRefs/CU:	-0.2020	0.294

The tests suggest neither length of text, nor density of new introductions per clause unit, correlates significantly with rates of RNS. Although the weak negative correlation of New Referents with RNS indicated in Fig. 2 points in the expected direction of the hypothesis, it does not reach significance.

Turning to the non-linguistic factors of familiarity, age, and gender, it likewise turns out that none of them appear to impact significantly on rates of NRS. Figure 3 provides the results based on the division into two groups, familiar and non-familiar, and Figure 4 the results according to speaker gender.

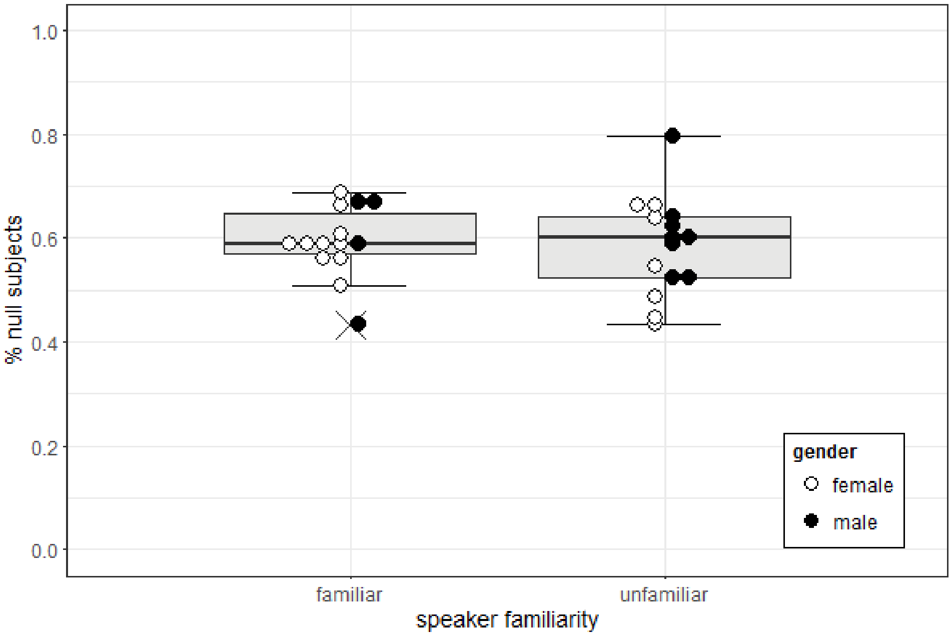

Figure 3: RNS according to speaker familiarity

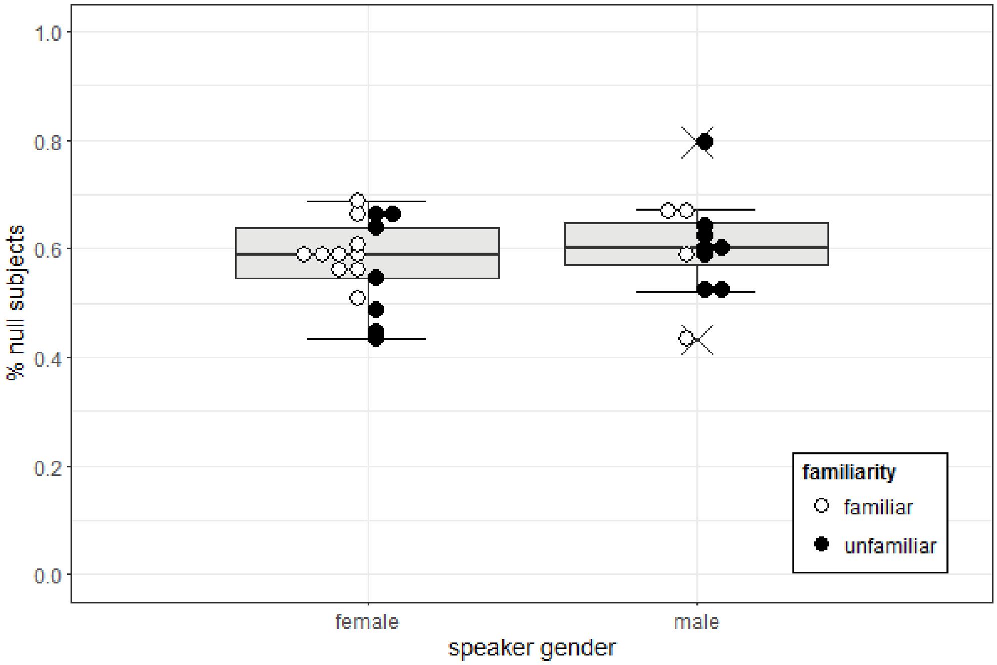

Figure 4: RNS according to speaker gender

The box-plot in Fig. 3 suggests that unfamiliarity leads to greater range of values than familiarity, but the overall mean of both the familiar and unfamiliar groups is similar, and the ANOVA test (Table 2) reveals no significant effect of speaker familiarity. Likewise gender does not appear to make an obvious difference. Age was also tested, but given the generally homogenous age grouping in the sample (the speakers' ages ranged from 20 to 39), age was not expected to be significant, and did not turn out to be (Pearson test for age: r=0.0644, p=0.740). The results of an ANOVA on all five variables is provided in Table 2.

Table 2: ANOVA of linguistic and non-linguistic factors

Factor	F	p
RNS | length	F(1, 27) = 0.403	p = 0.531
RNS | newR/CU	F(1, 27) = 1.148	p = 0.293
RNS | familiarity	F(1, 27) = 0.081	p = 0.779
RNS | gender	F(1, 27) = 0.726	p = 0.402
RNS | age	F(1, 27) = 0.113	p = 0.740

Based on these data, our conclusion is that rates of null subjects is remarkably stable across all speakers, regardless of age, gender, or degree of familiarity among the interlocutors. Thus we find no support for the hypothesis that speaker familiarity has an effect on RNS.

6 Discussion

Perhaps the most striking feature of the results is, disregarding for a moment two outlier values,[12] the stability of the RNS value across the sample as a whole (see e.g. the Standard Deviation (SD) value for RNS in Appendix 1). Whether our results generalize to other experimental settings remains an open question; they may be specific to Persian, or specific to the task, or simply reflect small sample size. However, our results actually appear well in line with the findings from research on spoken language registers summarized in Biber & Conrad (2009:261). Commenting on the results of several decades of research on variation across spoken registers, the authors note that "speech is highly constrained in its typical linguistic characteristics". Although written language displays considerable cross-register variation, "all spoken texts are surprisingly similar linguistically, regardless of communicative purpose (excluding scripted or memorized texts)." These conclusions may seem at odds with decades of research in the Labovian tradition of variationist sociolinguistics, which has sought to emphasize socially-determined variation in speech, but there is an important difference: most research in the variationist sociolinguistics paradigm continues to focus on phonology, rather than syntax (for example the phonological realization of the -ING suffix of English verbs has remained a "staple of sociolinguists" since the 1950's, (Hazen 2006)). Thus although we find it highly plausible that a social variable such as speaker familiarity would be reflected in phonological variation, or lexical choices, or perhaps intonation contours, it seems equally probable that syntactic features of discourse would be relatively stable, reflecting general cognitive constraints on short-term memory and instantiated through deeply entrenched and routinized patterns of delivery, mediated by language-specific morphosyntactic configurations. Biber & Conrad (2009) repeatedly point to the primacy of content and genre in determining variation in syntax. If this is indeed correct, then we would expect to find little variation across a sample of spoken texts of comparable content, regardless of setting. This prediction is borne out by our Persian data, where content was held fairly constant across all speakers.

7 Conclusions

Our investigation took up the challenge of investigating the factors that may impact on rates of null subjects in colloquial spoken Persian. We focussed on a possible impact of speaker familiarity, hypothesizing that greater familiarity among the interlocutors may lead to higher rates of null subjects in their speech, because familiar speakers can rely on a broader expanse of "common ground" (Matić et al 2014), and hence afford to be less explicit. Our investigation found little support for this idea, however. Rates of referential null subjects in spoken Persian instead appear to be relatively stable, and did not significantly correlate with speaker familiarity, or with the factors of gender and age. These findings are consistent with research on morphosyntactic variation in spoken language (Biber & Conrad 2009), which points to a high degree of cross-register homogeneity in spoken language, with the main determinants of variation being content and genre. The latter were held constant in our experimental design, which may help explain the overall lack of variation. However, we note that our data is almost entirely in the third person; dialogical data, involving first and second person forms, may pattern differently; this deserves further research.

[12] The two outliers are the speakers g1_m_13, with the unusually low RNS of 0.433, and g2_m_13 (RNS=0.796). We are unable to identify any biographic factors (e.g. bilingualism in another language) that might explain these extreme values.

Finally, we consider our research as an initial step towards an empirical and usage-based approach to syntactic variation in spoken Persian. Recently, corpus-based studies have opened up promising avenues for issues such as word-order variation in Persian (e.g. Faghiri et al 2014), and we expect that these developments will gather momentum in coming years. However, there is a considerable gap between written and spoken Persian, and as yet, most researchers interested in usage-based, as opposed to formalist, analyses of Persian have concentrated on the written language as their object of study (e.g. Roberts 2014), or on 'scripted spoken language', as in the film dialogues investigated in Vafaeian (2018). But with the exception of Frommer (1981) and Saeli and Miller (2018), there is very little empirical research on spontaneous colloquial spoken Persian. Our research is thus a modest attempt to develop corpus-building standards and methodologies for the future study of spoken Persian.

Appendix A: Raw figures from the experimental data, all speakers

speaker	familiarity	gender	age	CU's	RNS	NewRef	NewRef/CU
g1-f-01	familiar	female	39	47	0.558	15	0.319
g1-f-02	familiar	female	29	54	0.608	16	0.296
g1-m-03	familiar	male	22	17	0.588	8	0.471
g1-m-04	familiar	male	25	61	0.673	12	0.197
g1-f-05	familiar	female	26	60	0.510	17	0.283
g1-m-06	familiar	male	32	22	0.667	9	0.409
g1-f-07	familiar	female	25	38	0.588	11	0.289
g1-f-08	familiar	female	25	25	0.565	11	0.440
g1-f-09	familiar	female	25	100	0.663	23	0.230
g1-f-10	familiar	female	31	83	0.688	20	0.241
g1-f-11	familiar	female	33	60	0.593	13	0.217
g1-f-12	familiar	female	33	49	0.591	14	0.286
g1-m-13	familiar	male	35	69	0.433	17	0.246
g1-f-14	familiar	female	29	99	0.585	22	0.222
g2-f-01	unfamiliar	female	20	58	0.446	20	0.345
g2-f-02	unfamiliar	female	20	44	0.486	16	0.364
g2-f-03	unfamiliar	female	20	40	0.667	14	0.350
g2-f-04	unfamiliar	female	20	25	0.435	11	0.440
g2-f-05	unfamiliar	female	21	26	0.640	13	0.500
g2-f-06	unfamiliar	female	38	56	0.660	13	0.232
g2-f-07	unfamiliar	female	33	51	0.545	17	0.333
g2-m-08	unfamiliar	male	20	49	0.522	15	0.306
g2-m-09	unfamiliar	male	22	42	0.600	13	0.310
g2-m-10	unfamiliar	male	20	41	0.641	16	0.390
g2-m-11	unfamiliar	male	25	25	0.524	10	0.400
g2-m-12	unfamiliar	male	20	40	0.600	15	0.375
g2-m-13	unfamiliar	male	20	52	0.796	15	0.288
g2-m-14	unfamiliar	male	20	36	0.588	13	0.361
g2-m-15	unfamiliar	male	27	48	0.622	13	0.271
MEAN			26.03	48.86	0.589	14.55	0.325
SD			6.00	20.54	0.082	3.61	0.081

References

Alfaraz, Gabriela. 2015. Variation of Overt and Null Subject Pronouns in the Spanish of Santo Domingo. In Carvalho, Ana M., Rafael Orozco, Naomi Lapidus Shin (eds.) *Subject Pronoun Expression in Spanish: A Cross-Dialectal Perspective*, 5-18. Georgetown: Georgetown University Press.

Bell, Alan. 2006. Speech accomodation theory and audience design. In Brown, K. (ed.) *Encyclopedia of language and linguistics*. Boston: Elsevier, 648-651.

Biber, Douglas & Susan Conrad. 2009. *Register, genre, and style*. Cambridge: Cambridge University Press.

Bickel, Balthasar. 2003. Referential density in discourse and syntactic typology. *Language* 79, 708-736.

Bickel, Balthasar. 2011. Putting variation center-stage: Beyond "language" (or "dialect") as the basic data unit in language typology (and elsewhere). Keynote lecture at the conference: *Variation and typology: new trends in syntactic research*. The Linguistic Association of Finland, Helsinki, Aug. 25-27, 2011.

Camacho, José. 2013. *Null subjects*. Cambridge: Cambridge University Press.

Carvalho, Ana M., Rafael Orozco, Naomi Lapidus Shin (eds.). 2015. *Subject Pronoun Expression in Spanish: A Cross-Dialectal Perspective*. Georgetown: Georgetown University Press.

Chafe, Wallace. 1980. *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production*. Norwood, N.J.: Ablex,

Comrie, Bernard. 1989. *Language universals and linguistic typology. Syntax and morphology*. Chicago: University of Chicago Press.

Du Bois, John. 1987. The discourse basis of ergativity. *Language* 63(4):805–855.

Faghiri, Pegah, Samvelian, Pollet, & Hemforth, Barbara. (2014). Accessibility and Word Order: The Case of Ditransitive Constructions in Persian. In Stefan Müller (Ed.): *Proceedings of the 21st International Conference on Head-Driven Phrase Structure Grammar*, University at Buffalo, 217–237. Stanford, CA: CSLI Publications.

Falk, Yehuda. 2001. *Lexical-Functional Grammar. An introduction to parallel constraint-based syntax*. Stanford: CSLI Publications.

Farrell, Patrick. 2005. *Grammatical relations*. Oxford: Oxford University Press.

Frommer, Paul. 1981. *Post-verbal phenomena in colloquial Persian syntax*. PhD thesis, University of Southern California.

Ghomeshi, Jila. Forthcoming. Other approaches to syntax. In Anousha Sedighi and Pouneh Shabani-Jadidi (eds.) *The Oxford Handbook of Persian Linguistics*. Oxford: Oxford University Press.

Haig, Geoffrey. 2008. *Alignment change in Iranian languages: A Construction Grammar approach*. Berlin: Mouton.

Haig, Geoffrey. Under review. The pronoun-to-agreement cycle in Iranian: subjects do, objects don't.

Haig, Geoffrey and Stefan Schnell. Multi-CAST (Multilingual Corpus of Annotated Spoken Texts). URL https://lac.uni-koeln.de/multicast/ [accessed 15.01.2015].

Haig, Geoffrey and Stefan Schnell. 2014. Annotations using GRAID (Grammatical Relations and Animacy in Discourse), version 7.0. Available at: https://lac.uni-koeln.de/corpora/Multi-CAST/guidelines/Annotations/GRAID-Manual-7.0.pdf.

Haig, Geoffrey, Stefan Schnell and Claudia Wegener. 2011. Comparing corpora from endangered languages: Explorations in language typology based on original texts. In Geoffrey Haig et al. (eds.) *Documenting endangered languages: Achievements and perspective*s, 55-86. Berlin: Mouton.

Haig, Geoffrey & Stefan Schnell. 2016. The discourse basis of ergativity revisited. *Language* 92(3): 591-618.

Hazen K (2006), IN/ING Variable. In Keith Brown, (Editor-in-Chief) *Encyclopedia of Language & Linguistics, Second Edition*, volume 5, pp. 581-584. Oxford: Elsevier.

Holmberg, Anders. 2009. Null subject parameters. In Biberauer, Theresa & Anders Holmberg & Michelle Sheehan (eds.) *Parametric variation: Null subjects in Minimalist Theory*, 88-124. Cambridge: CUP.

Jahangiri, Nader. 1980. *A sociolinguistic study of Tehrani Persian*. PhD thesis, University College, London.

Keenan, Edward. 1976. Towards a universal definition of subject. In Li, Charles (ed.) *Subject and topic*, 303-333. New York: Academic.

Mahootian, Shahrzad and Lewis Gebhardt. 2018. Revisiting the status of -eš in Persian. In Alireza Korangy and Corey Miller (eds.), *Trends in Iranian and Persian Linguistics*, 263-276. Berlin: De Gruyter.

Matić, Dejan, Rik van Gijn and Robert D. Van Valin. 2014. Information structure and reference tracking in complex sentences: An overview. In Gijn, Rik van, Jeremy Hammond, Dejan Matić, Saskia van Putten and Ana Vilacy Galucio (eds.) *Information structure and reference tracking in complex sentences*, 1-41. Amsterdam: Benjamins.

Meyerhoff, Miriam. 2011. *Introducing sociolinguistics*. London: Routledge

Neeleman, Ad & Kriszta Szendröi. 2008. Case morphology and radical pro-drop. In Biberauer, Theresa (ed.) *The limits of syntactic variation*, 331-348. Amsterdam: Benjamins.

Orozco, Rafael. 2015. Pronominal Variation in Colombian Costeño Spanish. In Carvalho, Ana M., Rafael Orozco, Naomi Lapidus Shin (eds.) *Subject Pronoun Expression in Spanish: A Cross-Dialectal Perspective*, 19-40. Georgetown: Georgetown University Press.

Perlmutter, David. 1971. *Deep and surface constraints in syntax*. New York: Holt, Rinehart and Winston.

Peškova, Andrea. 2013. Experimenting with *pro-drop* in Spanish. *SKY Journal of Linguistics 26 (2013), 117–149*

Radford, Andrew. 2004. *Minimalist syntax. Exploring the structure of English*. Cambridge: Cambridge University Press.

Rasekh, Mohammad. 2014. Persian clitics: doubling and agreement. *Journal of Modern Languages* 24(1). 16–33.

Roberts, John. 2014. *A study of Persian discourse structure*. Uppsala: Acta Universitatis Upsaliensis.

Saeli, Hooman and Corey Miller. 2018. Some linguistic indicators of sociocultural formality in Persian. In Alireza Korangy and Corey Miller (eds.), *Trends in Iranian and Persian Linguistics*, 163-182. Berlin: De Gruyter.

Sato, Yosuke and Simin Karimi. 2016. Subject-object asymmetries in Persian argument ellipsis and the anti-agreement theory. *Glossa: a journal of general linguistics* 1(1): 8. 1–31.

Schiborr, Nils N. 2016a. English. In Haig, Geoffrey & Schnell, Stefan (eds.), Multi-CAST (Multilingual Corpus of Annotated Spoken Texts). (https://lac.uni-koeln.de/en/multicast-english/) [accessed 2016-02-26.]

Sedighi, Anousha. 2010. *Agreement Restrictions in Persian,* Leiden: Leiden University Press.

Stoll, Sabine & Balthasar Bickel. 2009. How deep are differences in referential density? In Lieven, E., J. Guo, N. Budwig, S. Ervin-Tripp, K. Nakamura, & Ş. Özçalişkan (eds.) *Crosslinguistic Approaches* *to the Psychology of Language: Research in the Traditions of Dan Slobin*, 543-555. London: Psychology Press.

Vafaeian, Ghazaleh. 2018. *Progressives in use and contact: A descriptive, areal and typological study with special focus on selected Iranian languages*. PhD thesis, Stockholm University, Faculty of Humanities, Department of Linguistics.