POLISH NULL SUBJECTS:

ENGLISH INFLUENCE ON HERITAGE POLISH IN TORONTO

by Joanna Chociej

Generals Paper I
Department of Linguistics
University of Toronto
March 2011

Supervisor: Naomi Nagy
Second Reader: Sali Tagliamonte
Third Reader: Alexei Kochetov

ABSTRACT

The present paper examines the variable use of null subject pronouns by Heritage Polish speakers living in the Greater Toronto Area. In contrast to the Polish use pattern, subject pronouns are close to categorically overt in English, the language of majority in the GTA. Thus, aside from documenting the grammar of null subject use in Polish, this paper investigates the extent to which the majority language can affect the grammar of a minority language.

On the basis of interview data from 15 Polish speakers, I examine the extent of English influence in relation to social and ethnic orientation factors. Speakers were divided into three groups: those who have lived in Poland their whole lives (homeland speakers); those who immigrated to Canada as adults (first generation heritage speakers); and those who were born in Canada or immigrated as very young children (second generation heritage speakers). The three groups were compared with respect to their overall rate of null pronouns, as well as the significant linguistic factors affecting their use.

Results show that overt subject pronouns are gaining frequency with each subsequent immigrant generation; that is, speakers who have greater contact with English tend to use typically English constructions more frequently. Heritage speakers also exhibit grammatical reanalysis: when compared to the grammar of homeland speakers, certain linguistic factors exhibit less significance while others exhibit more significance. Furthermore, while a speaker's null subject rate was not found to correlate to his/her ethnic orientation, correlations to other social and attitudinal factors such as the degree of language mixing and the amount of linguistic confidence were found.

Outside of contact linguistics, the analyses contained in this paper also give insight into the grammar of null subject use. When compared to previous variationist studies, the Polish data provide further evidence for the existence of universal patterns; most languages find subject continuity and type of person+number agreement to be significant factors in predicting null subject pronouns. Finally, as the first variationist study of Polish null subject use, this paper disproves the popular assumption that overt subject pronouns in Polish are essentially emphatic.

TABLE OF CONTENTS

1.0 INTRODUCTION

Contact induced language change has been gaining a lot of attention recently, both from the point of view of grammaticalization processes (Heine & Kuteva 2005) as well as the social factors involved. Heritage languages offer a unique perspective on the process: due to the degree of contact involved and the bilingualism of most heritage language speakers, changes are likely to occur more quickly. According to Trudgill (1983, cited in Heine & Kuteva 2005), "other things being equal (such as attitudinal factors, for instance), varieties whose speakers have frequent contact with speakers of other varieties will change more than varieties whose speakers have infrequent external contact".

Furthermore, as pointed out by Trudgill (*ibid*.), attitudinal factors may affect the degree of change. Heritage language speakers often have a choice between two different ethnic identities: one linked to the majority language and one linked to the minority language. As with any other social variable, the degree to which a speaker employs a particular feature may be an assertion of their identity. This may be true regardless of whether or not the speaker is conscious of his/her use of it. In the case of heritage language communities in Toronto, speakers who identify themselves as Canadian may choose to use English grammatical features when speaking their heritage language. In this paper, I investigate the relationship between a morphosyntactic variable in Polish (which has virtually categorical behaviour in English) and a range of social factors relating to degrees of language contact to quantitatively examine the connection between ethnicity and language variation. The results of this investigation suggest that while some attitudinal factors do play a role in resisting or facilitating grammatical change, overall ethnic orientation is not one of these factors.

1.1 Linguistic variable

Polish speakers form a significant community in Toronto. According to the 2006 census, 80,000 Toronto residents reported speaking Polish at home while 1.6% of all Greater Toronto Area (GTA) residents listed Polish as their mother tongue, making Polish one of the top ten mother tongues (after English). The Polish community has established community centres, churches, and heritage language programs. Additionally, Polish community members produce Polish-language media such as local newspapers, and television and radio programs. This social dynamic makes Polish a useful language to investigate in order to learn about language contact and dialect development.

To that end, a variable which is used to address the effects of language contact must have divergent usage patterns in the two languages. For heritage Polish speakers in Toronto, pro-drop is just such a variable. Also known as null subject variation[1], it is one of many linguistic features that sets English and Polish apart. As the example below shows, syntactic environments that require an overt pronoun in English (1a-b) do not require an overt pronoun in Polish (1c-d).

(1) a. Lately **we** studied a lot.

b. *Lately studied a lot.

c. **My** dużo ostatnio uczyliśmy się.
we much lately learn.1PL REFL

d. Dużo ostatnio uczyliśmy się.
much lately learn.1PL REFL

Within Polish grammar, a null subject is called *podmiot domyślny*, that is, an "assumed subject". This terminology is indicative of the attitude Polish grammarians and linguists have

[1] As there is debate about the existence of a pro-drop parameter and the appropriateness of the term "pro-drop" (see Mesthrie & Bhatt 2008), I will use the more neutral term "null subject" throughout this paper.

towards the phenomenon. As Gruszczyński & Bralczyk's (2002) dictionary of Polish grammar states, "sometimes the subject is unexpressed in the sentence, yet we can 'assume' what the subject is without difficulty based on the personal agreement on the verb or based on a sentence in the context" (p.190—my translation). While many grammars discuss the existence of null subjects in Polish, none to my knowledge talk about the degree to which Polish speakers employ it, and few talk about the contexts in which they appear. This is the first study to document the rate of null subject use among native speakers of Polish, as well as the linguistic variables that constrain its use.

1.2 Goals

The goals of this study are twofold. First, it sets out to define the grammar of null subject variation in Polish, and to determine whether any changes are taking place in the heritage language environment. To that end, it will answer the following questions:

1. What is the degree to which Polish speakers employ null subjects?
2. What are the factors that condition the variation between null and overt subject use?
3. Do heritage speakers of different generations employ null subjects to the same extent, and are they in control of the same conditioning factors as homeland speakers?

The secondary goals of this study concern themselves particularly with how the results can be generally applied to contact linguistics. Aside from helping determine which aspects of grammar are susceptible to majority language influence, the examination of ethnic orientation may help indicate which social factors play a role in resisting or facilitating that influence.

1.3 Overview

In the following sections I begin by describing the corpus which gave rise to the data used in this study (§2.1), and by circumscribing the variable context (§2.2). I move on to give an overview of existing literature on the topic (§3.1), and follow that with a critical discussion of how that literature compares with the current study (§3.2) and the hypotheses that the literature leads me to make (§3.3). I then outline the coding schema for the linguistic and social factors used in the current study (§4.1), and outline how each factor is expected to correlate to null subject use (§4.2).

Finally, I document the results of the analyses performed on the data (§5), including the distribution of the data (§5.1), the multivariate analyses of the contribution of social and linguistic factors (§5.2 and §5.3), and the correlation of null subject rate to social measures (§5.4). The implications of these findings are discussed in light of current linguistic and sociolinguistic knowledge (§6.1), as well as with an eye towards future linguistic research (§6.2). The paper concludes with an overview of the main findings (§7). Briefly, this study shows that the constraints behind null subject use are more complex than popular assumption would have us believe; the grammar of a heritage language changes with each subsequent immigrant generation; and external factors other than contact with English play a greater role in the degree and direction of change.

2.0 DATA AND METHODS

2.1 Data

2.1.1 Linguistic corpus

The primary data in this study is taken from recorded speech of 13 Polish speakers living in the Greater Toronto Area. They were recruited through social networking (family and friends of friends), as well as in response to an advertisement distributed by the Polish Students Association at the University of Toronto. Most interviews took place between June and August 2010, and the length of each interview was approximately one hour. Therefore, each speaker was fluent enough to sustain a fairly long conversation with a native Polish speaker. Two shorter interviews that were conducted in April 2009 as part of a pilot project were also included in the corpus.

The heritage speakers were divided into two groups based on their generation since immigration. There were six first-generation speakers who were born in Poland, moved to Canada as adults, and had resided in Canada for at least 15 years before the interview. In contrast, the seven second-generation speakers were either born in Canada or immigrated before the age of 5. Most of these speakers reside or grew up in western or southwestern parts of Toronto (see Figure A-1 in Appendix A), while they or their parents originate from southern or eastern Poland (see Figure A-2 in Appendix A).

Since, as far as I know, no sociolinguistic corpus of homeland Polish exists, two homeland speakers ("generation zero") were interviewed in addition to heritage speakers. These speakers were born in Poland and had been living there their whole lives. Their interviews were conducted at a time when they were both visiting Canada. This was their first trip outside of Poland, and they had been in Canada less than one month at the time of interview. Their data

facilitated comparisons aimed at detecting any differences between the speech of heritage speakers and that spoken in the homeland.

At least 20 minutes of an interview was transcribed for each speaker. Tokens were extracted from the beginning of the interview and coded starting with the second token (to allow for subject continuity data). At least 50 tokens from each speaker were included in the study. Since the homeland speaker group comprises only two speakers, twice as many tokens from each of these speakers were included in the analysis. The homeland speakers represent both younger (16 year-old) and older (44 year-old) speakers. However, they are not representative of the whole Polish population with respect to sex (both female) or geography (both grew up and reside in the same town). Socio-economically, they are both middle-class; the same can be said of all heritage speakers included in this study.

Table 2.1 includes a summary of each speaker, identifying their generation, sex, age at the time of interview, number of tokens included in the analysis, immigration date, and date of interview. The immigration date refers to the year in which they or their parents immigrated to Canada. Two years are noted if the year in which they emigrated Poland differed from the year of their arrival in Canada.

Note that throughout this paper speakers are referred to by a code comprised of their generation, sex, and age. For example, the first speaker in Table 2-1 is “0F44”.

Table 2.1: Speaker summary

Generation	Sex	Age (at time of interview)	Number of Tokens	Immigration Date	Interview Date
0	F	44	97	--	July 2010
(Homeland)	F	16	100	--	July 2010
1	M	88	59	1939 → 1951	August 2010
	M	60	54	1989	June 2010
	F	56	54	1993	April 2009
	M	55	76	1984	August 2010
	M	44	85	1989	August 2010
	M	35	57	1992	April 2009
2	M	47	52	1947	August 2010
	M	29	53	1970s	July 2010
	F	23	54	1990	July 2010
	M	23	53	1986 → 1987	July 2010
	M	22	53	1992	July 2010
	M	21	54	1983	July 2010
	F	18	56	1986 → 1988	June 2010

2.1.2 Ethnic orientation data

Each interview focused on the speaker's language use, history, and attitudes towards Polish language and culture. This casual sociolinguistic interview style allowed for collection of both linguistic and social data since social factors may influence the linguistic variable. A speaker's involvement in the community may be indicative of both their ethnic identity as well as the amount of contact they have with either Polish or English speakers.

The questionnaire that formed the basis of the interview was identical to that used by Nagy *et al.* (2010) and Hoffman & Walker (2010) in their measures of ethnic orientation in heritage and English speakers in Toronto. The original questionnaire was based on Keefe & Padilla (1987) and comprised 37 different questions. These questions were then reanalyzed and pared down to just 15 measures in the current study (see §4.1.2). The original questionnaire and speakers' responses can be found in Appendix B and C, respectively.

2.2 Variable context

2.2.1 Envelope of variation

As the example below shows, each personal subject pronoun in Polish has two optional variants: overt and null.

(2) *i ocenę w ogóle* ***Ø*** *miałam taką kiepską bo* ***ja*** *nie przeczytałam większość tych rzeczy*
"and [**Ø-I**] had a generally bad mark because **I** didn't read most of these things"
(2F18)

As per Otheguy *et al.* (2007:775), the envelope of variation includes all finite clauses with a null subject pronoun in an environment where overt pronouns are also likely to occur as well as all finite clauses with an overt subject pronoun in an environment where null pronouns are also likely to occur. Accordingly, token extraction was limited to tensed verbs, since non-finite verbs (including infinitives, participles, etc.) are inherently subject-less. The envelope was then narrowed down to tensed verbs which did not have a nominal subject, since pronoun subjects can only occur in contexts where a nominal subject is not present, and null forms are assumed to be variants of pronouns rather than nouns. Since Polish is a scrambling language, this included nominative subjects that occur either before or after the verb. Any contexts within this envelope where no variation between overt and null subjects appeared were then excluded (see for instance §4.1.1).

2.2.2 Exceptional distributions

A number of verbal contexts have exceptional distributions with respect to subject pronouns, and were therefore excluded from the study.

Conjoined subjects

In Polish, pronouns can be conjoined with other nouns either by is using the word *i* 'and' (3a) or by using the word *z* 'with' (3c). In environments where the coordination is made with *i* 'and', the pronoun must be overt (3b); if the coordination is made with *z* 'with', the pronoun may be either overt or null (3d).

(3) a. ***ja*** *i brat*
I and brother.$_{NOM}$
b. ****Ø*** *i brat*
and brother.$_{NOM}$
c. ***ja*** *z bratem*
I with brother.$_{INST}$
d. ***Ø*** *z bratem*
with brother.$_{INST}$

Furthermore, while a verb that has (3a) as its subject must have plural agreement, a verb that has (3c or d) as its subject can have either plural or singular agreement. Due to this exceptional distribution, all conjoined subject tokens were excluded from the study.

Interjections

Interjections are used stylistically and are often unrelated to the discourse topic. They are usually short, and as such tend to appear without subject pronouns in Polish (4).

(4) a. *wiesz* 'you know'
b. *rozumiesz* 'understand?'
c. *że tak powiem* 'I'd say'
d. *...myśle* '...I think'

Due to their stylistic function, they tend to be treated as set phrases that may be employed differently depending on the speaker. For instance, while most speakers say *nie wiem* to mean "I

dunno", other speakers may use other variants (5b-c). It is important to know that speakers who use other variants always use those variants to express this type of interjection.

(5) a. *nie* *wiem*
NEG know.1SG
b. *ja* *nie* *wiem* (1M55)
I NEG know.1SG
c. *czy* *ja* *wiem* (1M60)
Q I know.1SG

However, not all instances of these phrases are interjections. Therefore, while interjections of the type seen in (6a) were excluded from the study, cases where the phrase was a subject plus main verb (6b) were included.

(6) a. *ja* *jakoś,* *nie* *wiem,* *nie* *oddziałuje*
I somehow NEG know.1SG NEG take-part.1SG
"I somehow (I don't know) don't take part"

b. *ale* *jak* *dokładnie,* *to* *nie* *wiem*
but how exactly then NEG know.1SG
"but how exactly, I don't know"

(2M23)

Expletive constructions

Expletive constructions in Polish are inherently subject-less, meaning that there is neither an overt subject nor personal subject agreement on the verb (Gruszczyński & Bralczyk 2002:190-191 'zdanie bezpodmiotowe'). In such sentences, the verb has generic agreement (7a), person-less agreement ending in *–no –to* (7b), or a reflexive-type construction (7c).

(7) a. *Trzeba zrozumieć jej sytuację*
need understand.$_{INF}$ her situation
"There's a need to understand her situation"

b. *Niepotrzebnie zaproszono was tutaj*
needlessly invite.GEN you.$_{2PL.ACC}$ here
"You were invited here needlessly."

c. *Najefektywniej pracuje się z rana*
most-effectively work.$_{3SG}$ REFL in morning
"People work most effectively in the morning."

Because overt subject pronouns are unlikely to occur in this environment, these constructions were excluded from the study.

Non-personal pronouns

To "this /it" is a pronominal subject that also exhibits variability, as seen in (8).

(8) *chyba miał więcej wpływu od dziadków. Może* ***Ø*** *było tak, że dziadkowie przyjechali tutaj, pomagali rodzicom przez jakiś czas. Może tak* ***to*** *było, bo S. ma inny język, inny ten polski, inny akcent.*
"maybe he had more influence from the grandparents. Maybe [**Ø-it**] was so, that the grandparents came here, helped the parents for some time. Maybe **it** was so, because S. has a different language, this different Polish, different accent."
(1M44)

However, its generic nature means that it likely has a different function and distribution than personal pronouns. *To* also presents difficulties when it comes to coding since the form of the pronoun is the same as a common clause marker (see §4.1.1). Therefore, this non-personal pronoun falls outside the scope of this paper.

Imperative constructions

Gruszczyński & Bralczyk (2002) include examples of imperative constructions under their entry on "null subjects" as opposed to "subject-less clauses", which suggests that, unlike their English counterparts, Polish imperative constructions can have overt subject pronouns. Accordingly, it is not expected that the overt pronoun rate in this context will be greater as a result of English influence. Due to this opposite dynamic, imperative constructions have been excluded.

Relative clauses

Relative clauses where the relative pronoun is coreferential with the subject of the verb are not variable with respect to null subjects—they are always null (9a) in both English and Polish. However, in cases where the relative pronoun is coreferential with the object of the verb, variation does exist (9b). Therefore, while the first type of relative clause has been excluded from analysis, the second type has been included.

(9) a. *bluzka* *która* *(*ona)* *teraz wisi* *na* *drzwiach*
shirt which she now hang.$_{3SG}$ on door
"the shirt which (*it) now hangs on the door"

b. *bluzka* *którą* *(ty)* *kupiłaś* *mi* *wczoraj*
shirt which.$_{ACC}$ you.$_{SG}$ buy.$_{2SG.PAST}$ me yesterday
"the shirt which you bought me yesterday"

Ambiguous contexts

Ambiguous contexts include false starts, corrections, repetitions, verbal lists, and meta-linguistic commentary. In cases where two tensed verbs are uttered one after another, and provided that the first verb has been uttered completely and not cut off mid-word, only the first

of these verbs counts as a token (underlined in 10a). This is because subsequent verbs in this type of sequence can be interpreted as conjoined VPs, in which case they cannot have their own subject. Similarly, repetition of the same verb at the end of a clause can be interpreted as stylistic emphasis, and as such is also excluded from the data. For instance, in the speech sequence found in (10b) only two tokens (underlined) were counted.

(10) a. *się tam <u>urodziłem</u> wychowałem* (1M55)
REFL there born.$_{1SG.PAST}$ raised$_{1SG.PAST}$
"I was born there, raised"

b. *<u>nie mam</u> rodziny nie mam nie mam*
NEG have family NEG have NEG have
za dużo <u>nie mam</u> do czego wracać (1M55)
too much NEG have to what return.$_{INF}$
"I don't have family. No, no. I don't have much to return to."

Finally, any type of quoted meta-linguistic commentary was also excluded. For example, when talking about past versus present tense, (1M88) produced a number of examples of conjugated verbs (11), but only the quotative *powie* 'say' is included in the data.

(11) *on <u>powie</u> "ja wczoraj jem"... "jem to" nie "jadlem"*
he say.$_{FUT}$ I yesterday eat.$_{PRES}$ eat.$_{PRES}$ this NEG eat.$_{PAST}$
"he will say, 'I eat yesterday'... 'eat this', not 'ate'."

3.0 SITUATING THE LINGUISTIC VARIABLE

The current section situates the present study within the existing body of literature on null subject pronouns and contact linguistics. I begin with an objective overview of literature, relating the claims and findings or lack of findings therein (§3.1). Subsequently, I comment on the literature by discussing its relation to the current study and what the current study will contribute to knowledge on the subject (§3.2). Finally, the conclusions reached in the literature are reiterated along with the hypotheses that will be tested in the current study (§3.3).

3.1 Literature Review

The literature review begins by focusing on the Polish language in particular. It gives an overview of existing sociolinguistic literature on the Polish language (§3.1.1) and moves on to describe the available information on the status of null subjects in Polish (§3.1.2). It then broadens its scope and discusses features of null subject languages in general (§3.1.3) and null subject languages in contact situations (§3.1.4). It concludes by detailing what is known about null subjects in English (§3.1.5) and about the process of grammaticalization resulting from language contact (§3.1.6).

3.1.1 Polish sociolinguistics

Most of the existing literature on Polish sociolinguistics is not relevant to the present study. Research on "contemporary" and "colloquial" Polish often aims to document lexical and phonetic phenomena (c.f. Bartmiński 2001; Dubisz 1997; Lubaś 2003), while glossing over syntactic and morphophonological features. The same can be said of dialect studies—both those

that focus on urban versus rural areas in Poland, as well as those that focus on the language of Poles living outside of Poland.

Lustanski's 2009b work is the most pertinent for the present study, as it offers a recent analysis of the language of Polish speakers living in the GTA. It is also the most extensive study in terms of grammatical features examined, with chapters on phonetics and phonology, morphology, lexicon and word-formation, and syntax. Her study includes data from 33 first-generation Polish immigrants who came into Canada as adults, and 182 second-generation young Polish Canadians who were either born in Canada or immigrated as very young children. Workers of the company MCM2001 in Toronto represented the first generation, while students from the ninth, tenth, and eleventh grade at the M. Copernicus Polish School in Mississauga represented the second generation.

In her chapter on syntax, Lustanski points out that Polish speakers in Toronto tend to overuse overt pronouns. This phenomenon can be partially attributed to the loss of subject agreement on the verb and the general "tendency to not inflect verbs" (p.158 – my translation). Lustanski goes on to further say that "if the sentence has a grammatical subject that is not null, then an infinitive form appears next to it" (p.158 – my translation).

(12) a. *oni gwizda-ć*
they whistle-$_{INF}$
"they whistle"

b. *oni gwizda-ją*
they whistle-$_{3PL}$
"they whistle"

However, Lustanski does not include any statistical support for her claims that overt pronoun subjects are overused even in cases where the verb does carry agreement. Like many

other sociolinguistic studies on Polish, her work lacks information about the degree to which any of the documented features are used. As such, it is impossible to say how much overt pronoun use is considered "overuse" in comparison to standard Polish.

3.1.2 Null subjects in Polish

Despite the lack of linguistic and sociolinguistic studies on null subjects in Polish, there are certain popular assumptions in the literature. However, many of these assumptions have up until now remained untested. One such assumption can be found in Lustanski's connection between overt pronouns and infinitive forms in §3.1.1 above. Along with other researchers, Lustanski posits that overt subject use is somehow related to lack or loss of subject agreement on the verb.

Polinsky's (1995) survey of six heritage languages concludes that, for speakers of null subject languages, inflated use of overt pronouns compensates for the loss of agreement. Her study examined speakers of Eastern Armenian, Lithuanian, Polish, Russian, Kabardian, and Tamil who do not speak these languages as their primary language and show significant attrition. Interestingly, these speakers, "who generally show a high rate of acceptance [of] any sequences (including ungrammatical and marginal ones), did not accept pro-drop examples elicited from native speakers; they corrected such examples by inserting overt pronouns or full NPs" (p.97). Furthermore, in her comparison of a number of linguistic features, Polinsky found a positive correlation (Spearman rank correlation coefficient = 0.885) between correct agreement and null subject use (p.113).

The assumed link between null subjects and verbal agreement is further reiterated by researchers such as Bondaruk (2001:35), who states that "Polish 'drops' unstressed pronominal subjects for all persons and numbers, as its verb forms carry a distinct inflectional ending in

every person and number." McShane (2009:201), in her comparison of null subject use in Polish and Russian, expands the agreement argument further by saying that "Polish requires it in more contexts than Russian—which might be a reflex of its richer verbal morphology in certain forms." For her, rich verbal morphology does not stop at person and number; rather, Polish employs null subjects more because it also inflects for gender in the past tense (endnote).

However, while it is easy to assume that null subject use is tied to verbal agreement, Lindseth (1998:2-3) points out that this relationship is less direct than we may be inclined to believe. He states that "on the one hand, there are languages like German, which seem to have what we would call rich agreement, but fail to allow referential null subjects, while on the other hand, there are languages like Japanese and Chinese, which are quite capable of omitting subject pronouns, despite their lack of agreement morphology." Therefore, at the very least, rich verbal agreement cannot be the only factor in null subject use.

Lindseth (1998:34) goes on to give what he considers to be typical characteristics of null-subject languages.

(13) a. Only null pronominals are stylistically unmarked.
b. Only null pronouns can function as bound variables.
c. Only null 3rd plural pronominal subjects can have arbitrary reference.

The first of these characteristics is the basis of another common assumption. Lindseth explains that "the use of overt pronouns is generally reserved for contrast, emphasis, or a change in discourse topic" (p.40). Frajzyngier & Shay (2003:156-157) support this observation by citing Meillet (1937) who "states explicitly that a clause with an independent subject pronoun in Latin has a different meaning from a similar clause without the pronoun". Therefore, Frajzyngier & Shay view subject pronouns in Polish as a means of coding switch reference and contrast, which

is independent of agreement marking on the verb. Similarly, Bondaruk (2001:37) takes the correlation between overt pronouns and emphasis even further by stating that "whenever [overt pronouns] are retained they bear heavy stress".

Nonetheless, Frajzyngier & Shay go on to explain that overt subject pronouns in Polish are further limited in their function as markers of switch reference. "In many languages, coreference with the closest preceding referent of the same gender, number, person, and case requires the least morphological marking" (Frajzyngier & Shay 2003:251). In the case of Polish, this marking involves only agreement on the verb. "This coding means is used even when another potential subject argument intervenes between the previous mention and the new mention, provided this argument has different features for gender, number, or person" (p.257). That is, overt subject pronouns are used to code switch reference only if the preceding subject has the same features for gender, number, and person.

Similarly, McShane's (2009) analysis of null and overt subjects in Polish and Russian found many reference-based limitations. McShane's approach follows the theory of Ontological Semantics, which necessitates that all her proposed constraints be sufficient to provide a non-native speaker or intelligent agent with practical knowledge of subject pronoun usage. According to McShane, while subject pronouns in Polish have a baseline null realization, certain factors can overrule that baseline. Most of these factors (summarized in 14) are semantic or pragmatic in nature, and arise from a potential for ambiguity (see McShane:109-118 for full elaboration).

(14)

a. The antecedent is insufficiently syntactically accessible to guarantee recoverability of the elided subject.
b. The third person subject of a subordinate clause is not coreferential with the subject of the matrix clause.

c. There is a shift in the subject or the agent/experience in sequential main clauses.
d. There is a shift in agent/experience between denominal and a subsequent tensed clause.
e. The antecedent is a rhematic subject as detected by word order.
f. The personal ending of the verb form is phonetically ambiguous.

Without going into too much detail about each of these factors, it is important to note that most of them (a-e) are in some way related to the particular syntactic position of the pronoun's antecedent. However, one factor (f) is morphophonological in nature. In particular, it refers to the fact that 3rd person present tense forms ending in *–e* /ɛ/ and 1st person present tense verb forms ending in *–ę* /ɛ̃/ sound identical in fast speech [ɛ]. According to McShane, this neutralization context found in common verbs such as *chcę* 'I want' "can lead to the use of a disambiguating subject pronoun" (p.115). This may be one of the reason behind McShane's observation that while the subject tends to be null in written Polish, there are considerably more overt subjects in spoken Polish than in the written language (p.107).

3.1.3 Variation in null subject languages

Many sociolinguistic analyses of null subject use have been conducted on other languages, suggesting a number of factors necessary to consider in the present analysis. The most popular language in these types of studies is Spanish. As Otheguy *et al.* (2007:772-773) summarizes, "the independent grammatical variables linked to the probability of occurrence of an overt pronoun include the person and number of the verb (Silva-Corvalán 1994), the verb tense (Silva-Corvalán 1982, 1997b), the type of clause where it appears (Morales 1997), the discourse status of the subject's referent [subject continuity] (Cameron 1996, Silva-Corvalán 1982, 1994), and the lexical content of the verb (Enríquez 1984, Silva-Corvalán 1994)." The lexical distinctions made by Otheguy *et al.* are: mental or estimative verb, stative verb, and external action verb;

while the clause type distinctions are: main clause, subordinate relative clause, subordinate arguments, other subordinate clauses, coordinate clauses, and other clauses. Morales (1997) found a greater probability of overt pronouns appearing in object relative clauses than in main clauses. Bayley & Pease-Alvarez (1997) examined the significance of surface ambiguity of inflected verb forms. Runs by verb type indicate that pronouns are likely to be expressed overtly with imperfect, conditional and subjunctive forms regardless of whether they preserve their morphological distinctness. Therefore, they argue that tense/aspect features are a better explanation for higher incidence of overt with imperfect/conditional/subjunctive than the functional compensation hypothesis.

If we compare the findings for Spanish with the findings for other null subject languages, such as Heritage Russian, Cantonese, and Italian, examined by Nagy *et al.* (2010) we find a number of trends. Most importantly, subject continuity is significant in all languages (Spanish, Russian, Cantonese, Italian), with verbs whose subjects have the same referent as the subject of the preceding verb occurring with more null subjects than verbs whose subjects have a different referent from the subject of the preceding verb. Also, while the effect within the person or number factor is different depending on the language, all languages do find one or both of them significant in some way. Person is significant for Russian and Cantonese; number is significant for Italian; and the combination of person and number is significant for Spanish. Other significant factors include tense (Spanish and Italian), and clause type (Spanish, Cantonese, and Russian). Additional factors were found significant depending on the particular grammar of the language involved (e.g. preverbal direct object for Italian; negation for second/third generation Russian). Most interestingly, independent social factors were found significant for Russian (speaker age group for all generations; speaker sex for second and third generations).

3.1.4 Null subject languages in contact

Many studies of heritage languages point to the degree of contact with the majority language as a significant factor in the degree of change. Most studies report that contact with English results in greater use of overt pronouns by speakers of null-subject languages (see Heine & Kuteva 2005:68-70 who cite Myers-Scotton 2002 for Spanish, Schmitt 2000 and Polinsky 1995 for Russian, Savić 1995 for Serbian, and Bolonyai 2000 for Hungarian). However, this is not always the case. Torres-Cacoulos & Travis (2010), in their review of studies done on Spanish in the United States, report both *increased* use of overt pronouns with greater contact with English (Lipski 1996, Montrul 2004, Toribio 2004, Otheguy *et al.* 2007), as well as a *decreased* use of overt pronouns with greater contact with English (Silva-Corvalan 1994, Bayley & Pease-Alvarez 1997, Travis 1997). In particular, Otheguy *et al.* (2007) found that the rate of overt pronouns in Spanish rose from 30% (for newcomers to New York) to 38% (for New York born and/or raised). This increase in rate was true for both speakers of "mainland" Spanish (a change of 36% to 42%) and "Caribbean" Spanish (a change of 24% to 33%). On the other hand, in their study of Mexican Spanish speakers living in California, Bayley & Pease-Alvarez (1997) found that English-dominant children are less likely to use overt pronouns than children born in Mexico. Speakers born in the US whose mothers were born in the US or immigrated at 10 or younger had an overt pronoun rate of just 16%, while speakers born in Mexico or those whose mothers immigrated at 15 or older had much higher overt pronoun rates (25% and 27% respectively).

Nonetheless, some studies show that contact with English has no effect on the null subject rate. Nagy *et al.* (2010) found that for Heritage Cantonese, Italian, and Russian generation since immigration (that is, whether the speakers immigrated to Canada as adults or whether they were

born there) was not a significant factor in the null subject rate across all languages studied. In response to these findings, Nagy *et al.* speculate that English contact greatly influences even the first-generation speakers. Furthermore, individual speakers showed no correlation between null subject rates and ethnic orientation score, which suggests that speakers do not use this particular variable to index ethnic orientation.

The idea that language use may not be an indication of ethnic orientation is somewhat borne out in Lustanski's (2009a) survey of Polish speakers in Toronto. When asked what Polish immigrants should do to maintain their "Polishness", only 38% of her respondents indicated "speak Polish" and only 16% indicated "teach children Polish" (p.48-49). Furthermore, when asked "What brings Polish people together?", both first and second generation speakers place "possibility of conversation in Polish" as third out of four possible responses (p.54-55). These results suggest that language and language maintenance is not a primary concern for Poles in Toronto.

However, the identity of the factors that affect the susceptibility of a speaker's linguistic features to language contact remains unknown. Niedzielski (1997) compared Polish speakers in France, United States, and New Zealand, to determine the correlation between language attitudes and linguistic maintenance in these Polish immigrant communities. According to him, "linguistic assimilation or resistance to assimilation depends to a great extent on cultural assimilation of the group of immigrants" (p.1767). His findings suggest a number of factors involved in first-language maintenance (15) (p.1777).

(15)

1. The amount of education the immigrants received in the first language and culture before joining the new culture and language.
2. The desire of the group to be identified as minority ethnic; either they are proud of it or they suffer no stigma because of it.

3. The degree with which they avoid using their first language with less than proficient people.
4. The homogeneity of the host society and the pressure it may exert on the minority group.
5. In addition, when the minority group is dialectally heterogeneous, it tends to standardize its first language.

The two factors on Niedzielski's list that are pertinent to the current study are #1 and #4. With respect to language education, Niedzielski cites the case of orphan children who were relocated from Poland to Auckland, New Zealand during the Second World War. He points out that language maintenance levels of Auckland Poles are high despite the fact that most of their education took place in special Polish community schools in New Zealand. This is further supported by Polinsky's (1995) research on heritage language speakers which found a positive correlation between the maintenance of a language and the time spent in the language community, despite the fact that there is no significant difference between those speakers who were born into a L1 community and those speakers who were born in the U.S., nor is there a difference between those who left the L1 community before age 7 and those who left after that age (p.117).

Elaborating on factor #4, Niedzielski contrasts the situation of Poles in France and Poles in the U.S. Polish speakers in France exhibit greater language maintenance despite profuse encouragement to speak French everywhere, particularly at work. On the other hand, Polish speakers in the U.S. exhibit less language maintenance despite having the option to speak their heritage language at work. This observation is related to Otheguy *et al.*'s (2007) (somewhat) surprising finding that there was a positive correlation between a high overt pronoun rate and using Spanish with workmates. In other words, while more Spanish use at home (with siblings

and parents) resulted in a lower overt pronoun rate, more Spanish use at work resulted in a higher overt pronoun rate.

3.1.5 Null subjects in English

Nonetheless, we cannot claim that null-subject heritage languages are moving towards English, if we do not know how null subjects behave in English. Harvie (1998) attempted to analyze this by examining data from 14 working-class 15-45 year-old native English speakers in Ottawa. Harvie coded for the following factors: subject type, clause type, negation, position in clause, switch reference, contrast, and turn position. However, only switch reference and position in clause were found significant, with same referents and initial position in clause having the most null subjects. While number was not significant, first person was least frequently null. Furthermore, separate runs for main and conjoined clause types suggest that they behave similarly with respect to null subjects.

However, Nagy *et al.*'s (2010) analysis of English found that a factor of subject continuity and conjunction combined is the only significant one. The factor weights for each context type show a continuum of most null to least null: same referent conjoined (.86) > same referent main (.53) > switch referent main (.34) > switch referent conjoined (.21). In general what this shows is that subjects with the same referent as the preceding subject have more null pronouns than subjects with a different referent than the preceding subject, and that this difference is greater in conjoined clauses than in main clauses.

3.1.6 Language contact and grammaticalization

In any study of language change, it is also important to distinguish contact-induced change from universal language shift and attrition. The first question is of course whether a drift towards a high frequency of overt pronouns is possible without a language contact environment. Lindseth (1998:128) argues that in order for this type of internal language change to happen, "null and overt pronouns in that particular language must have the same referential properties. Otherwise, they will not be in competition and the pro-drop character of the language will remain stable" (p.128). Since, according to Lindseth, "the overt pronoun and the zero form in [canonical null-subject languages] do not represent competing forms" (p.129), an internal language drift cannot occur.

On the other hand, in situations of language contact, the process of grammatical replication can lead to the reanalysis of overt pronoun function. According to Heine & Kuteva (2005), "grammatical replication has the effect that the replica language (R) acquires some new structure (Rx) on the model of another language (M) […] however, the new structure Rx is in most cases not entirely new; rather, it is built on some structure (Ry) that already existed in the replica language, and what replication then achieves is that it transforms Ry into Rx" (p.40). In essence, what replication does is take an existing minor use structure (Ry), which is optional and restricted in its occurrence to some specific grammatical meaning, and transforms it into a major use pattern (Rx) by expanding its function.

According to Heine & Kuteva (2005), the rise of a major use pattern in contact-induced replication follows the three chronological steps set out below (p.45):

(16) a. Frequency: An existing use pattern is used more frequently.
b. Context extension: It is used in new contexts.
c. Change in meaning: It may become associated with a new grammatical function.

When this process is applied to null-subject languages, overt subject pronouns, which originally served pragmatically defined functions (e.g. emphasis, switch reference), gain in frequency and increasingly lose their original function. Once they are used in new (i.e. non-pragmatic) contexts, they undergo context-induced reinterpretation, eventually assuming the function of regular subject markers. However, it is possible that the change in meaning may not necessarily lead to a completely English use pattern. As Johanson (2008:77) points out, "codes do not completely abandon their heritage and do not fuse with unrelated contact codes. No high-copying code seems to have turned into the Model Code it has copied extensively from."

Furthermore, Heine & Kuteva (2005) contrast contact-induced grammaticalization with attrition by saying that contact-induced grammaticalization "leads to an enrichment of the language concerned, in that new use patterns and grammatical categories are created on the model of another language", whereas in attrition "existing categories are simplified, merge with other categories, or are simply abandoned" (p.256). In other words, "speakers of dying languages tend to over generalize unmarked features at the expense of marked ones" (p.255). Therefore, since overt subject pronouns are marked, they are less likely to gain frequency in cases of attrition.

3.2 Commentary: existing literature and the current study

The weakest points of existing literature are the glaring gaps in the data. For instance, while Lustanski (2009a) summarizes that Heritage Polish speakers in Toronto overuse overt subject pronouns, she does not back up her claims with statistical information about overt subject pronouns in Heritage Polish versus Homeland Polish. Neither does she specify whether her observation is equally true for all or for just some of the heritage speakers she surveyed.

Similarly, Nagy *et al.*'s (2010) study of heritage Italian, Cantonese, and Russian did not find any difference between first generation and second generation heritage speakers. As they themselves admit, conclusions about whether English does or does not influence heritage languages cannot be drawn from their results since it is possible that exposure English may have affected both first generation and second generation speakers equally. That is, there needs to be information about how a particular variable behaves in the homeland variety, where English influence does not play a role, in order to determine whether heritage varieties have adopted English features or whether these features have been present in the language from the beginning. Therefore, the inclusion of data from two homeland speakers in the current study makes it the first of its kind to document the behaviour of a grammatical feature in both homeland and heritage varieties using the same methods.

Furthermore, there are additional major differences in speaker sample between the current study and previous analyses of Polish. One difference has to do with the homogeneity of other speaker corpora. For instance, in Lustanski's (2009a,b) data, all first generation speakers work at the same company, while all second generation speakers attend the same school. This homogeneity severely restricts her ability to differentiate her speakers based on social factors. Polinsky (1995), on the other hand, limits her data to "terminal" speakers who exhibit significant attrition. Specifically, they all cite English as their primary language, simplify case inflection, and simplify subject-verb agreement. In comparison, the speakers in the present study either cite Polish as their primary language or as being on equal terms with English. Most importantly, most of my speakers use correct inflection and agreement markers. Even the one speaker (2M47) who exhibits some attrition with respect to case inflection does not use ambiguous subject agreement. His most common error with respect to subject agreement involves

conjugating verbs with morpheme forms belonging to another inflection class. The example in (17a) and its standard Polish equivalent in (17b) shows the speaker using the first person singular suffix *–em* /ɛm/ instead of the equivalent *–ę* /ɛ̃/ (see §4.1.1 'Lexeme' for information about verb inflection classes).

(17) a. *tu ja myślem że my jesteśmy za otwarci do obce*
here I think.$_{1SG}$ that we are.$_{1PL}$ too open.$_{3PL}$ to other.$_{NOM}$
ludzie (2M47)
people.$_{NOM}$
"I think that we are too open to strangers here."

b. *tu ja myślę że my jesteśmy za bardzo otwarci do*
here I think.$_{1SG}$ that we are.$_{1PL}$ too much open.$_{3PL}$ to
obcych ludzi
other.$_{GEN}$ people.$_{GEN}$

Therefore, if any of my speakers exhibit greater overt pronoun use, it is not possible for me to argue that this is in any way due to loss of agreement on the verb.

Finally, certain factors discussed in the literature cannot be coded, either because they are impossible to judge, because they are irrelevant to this particular language, or because they fall outside the scope of this paper. For instance, McShane (2005) used exclamation points to judge sentential emphasis in written text, but admits it would be near impossible to judge sentential emphasis in speech. She further notes that, in speech, the presence of an overt pronoun does not mean that the sentence is emphatic. Therefore, although the popular belief is that overt subject pronouns in Polish indicate emphasis, it is beyond the scope of this project to code emphasis as a linguistic variable. Nonetheless, it is possible to partially examine emphasis by coding for certain types of inherently contrastive constructions. Within the present study, these include clauses headed by contrastive coordinating morphemes, as well as some cases of switch reference, negation, and interrogatives.

Furthermore, although subject continuity is unarguably an important factor in null subject use, employing syntactic distinctions such as those suggested by Frajzyngier & Shay (2003) and McShane (2005) falls outside the scope of this paper. Both articles detail a complicated system based on the multiple syntactic positions of possible antecendents. This approach to subject continuity would alone constitute a major research project. Therefore, for the purposes of the current study, a more simplified distinction of subject continuity is employed.

3.3 Hypotheses

The existing literature suggests a number of hypotheses regarding the linguistic and social factors that are significant for null subject use in heritage languages. Each of the hypotheses outlined in this section are based in the above literature review, and they will be revisited in §6.1 wherein their credibility will be discussed in light of the results and analyses presented in §5.

The first question to ask is whether or not grammatical change is expected in a heritage language environment, and if so what kind of change will it be. Heine & Kuteva's (2005) research provides clues towards the answer. First, they point out that in cases of attrition, the categories within a grammar are simplified due to the push to be more unmarked. Accordingly, provided that speakers retain subject agreement marking, they will continue to use null forms (i.e. less redundancy). On the other hand, situations of language contact would cause new categories to be created. Therefore, since the heritage language environment is a situation of language contact, I would expect increased use of overt subject pronouns with greater language contact.

Additionally, different patterns are expected depending on which stage of the grammaticalization process the speaker is at. Based on Heine & Kuteva (2005), I would first

expect a change in frequency, then a narrowing in the significance of pragmatic factors (i.e. subject continuity), then grammatical reanalysis where factors that are not used by homeland speakers have become significant in lieu of subject continuity. I would not expect more than a change in frequency in first generation speakers, since it would be unreasonable to expect a greater degree of grammaticalization within one generation. On the other hand, second generation speakers may take the grammaticalization process one step further and extend the overt pronoun context so that switch reference is no longer a significant factor.

However, Heine & Kuteva's (2005) hypothesis that the grammaticalization process diminishes the significance of pragmatic factors is complicated by Harvie's (1998) findings for English. Specifically, Harvie points out that subject continuity is a significant factor in English as well as null-subject languages, which in turn hints at the universal nature of the null subject phenomenon. If subject continuity is significant for English, it is possible that it will remain significant for heritage Polish. Nonetheless, it is expected that subject continuity will be significant for at least some speakers of Polish, not in the least because it has been found to be significant for every other null subject language studied (c.f. Otheguy *et al.* 2007; Bayley & Pease-Alvarez 1997; Nagy *et al.* 2010).

Other factors that were found to be significant for null subject languages include: grammatical person and number of the subject; tense/aspect of the verb; clause type; presence of negation; and lexical-semantic content of the verb (see §3.1.3). It may be that these factors are also significant for Polish, and so they will each be tested in the present study. Whereas the significant status of most of these factors depends on the language, grammatical subject type was found to be a significant factor for all languages. Different studies made different distinctions with respect to this constraint: some treated grammatical person separately from grammatical

number, while others used an intersection of person and number in their coding schemas. Nonetheless, all studies of null subject languages report finding either person or number or both as significant. Therefore, I expect that this constraint will be significant for Polish as well.

As discussed in §3.1.2, popular belief holds that null subjects are correlated with subject agreement on the verb, while overt subjects in null subject languages are correlated with emphasis. Even though all Polish finite verbs must bear number and person agreement, the agreement system of Polish distinguishes between tenses which mark for speaker gender and tenses which do not (§4.1.1). Therefore, I would expect the presence of gender agreement to coincide with null subject use in Polish.

One final factor claimed to be significant for Polish has to do with phonetic ambiguity of the verb form. McShane (2005) argues that Polish verb forms inflected with a morpheme that may be phonetically ambiguous with respect to the person it refers to will have more overt pronoun subjects. However, Bayley & Pease-Alvarez (1997) in their research on Spanish have found that tense/aspect features associated with phonetically ambiguous forms offer a better explanation for higher incidence of overt pronouns than the functional compensation hypothesis. That is to say, specific phonetically ambiguous forms are not any more likely to occur with overt pronouns than any other verb form inflected for the same tense/aspect feature. When applied to Polish, these findings suggest that either phonetic ambiguity in particular will be a significant factor for null subject use (based on McShane 2005), or that the significance will lie in tense/aspect features that contain phonetically ambiguous forms in general (based on Bayley & Pease-Alvarez 1997).

Finally, social factors affecting grammar at the individual level have yet to be identified. Many studies suggest that the amount of contact and the depth of ties that a speaker has with

either the heritage language and culture or the majority language and culture will have an effect (c.f. Otheguy *et al.* 2007; Bayley & Pease-Alvarez 1997). However, as Lustanski's (2009a) research showed, Polish ethnic identity may not be associated with the Polish language for all speakers. Therefore, ethnic orientation may not be a significant factor in null subject use. Furthermore, as Niedzielski (1997) points out, other factors may further resist or attract influence from the majority language. One of the factors he cites involves the amount of language education the immigrants received, and consequently the amount of awareness they may have about what constitutes "proper" Polish. Another factor on Niedzielski's list deals with the amount of external pressure that society exerts on the immigrants to assimilate, and consequently the number of environments where immigrants feel they can use both languages.

All the hypotheses outlined above regarding Polish null subject use, contact-induced language change, and ethno-linguistic social factors will be tested in the current study.

4.0 CODING AND ANALYSIS

A number of factors were coded in the present study to test the hypotheses outlined in §3.3. These include both linguistic factors (§4.1.1) as well as ethnic orientation factors (§4.1.2). The sections below outline the coding distinctions used in the present study (§4.1), and revise the hypotheses from §3.3 to make reference to particular factors within factor groups (§4.2).

4.1 Coding schema

4.1.1 Linguistic factors

Each token was coded for 9 linguistic factors: overt verbal morphology including tense, person, number, and gender; lexeme; negation; sentence type; subject continuity; and clause type.

Overt verbal morphology

In addition to being marked for tense, verbs appear with agreement morphology reflecting the person, number, and sometimes grammatical gender of the subject. The coding schema made four distinctions with respect to tense marking: present; past; future perfect; and future imperfect. Verbs marked for the conditional/subjunctive tense were also extracted and coded separately. However, after extracting 50 such tokens, their subjects were found to be categorically null, and all conditional/subjunctive tokens were ultimately excluded from any analysis.

All tensed verbs are inflected for person and number agreement with the subject. The shape of the morpheme depends not only on the tense of the verb, but also on its inflection class (c.f. 'Lexeme' below). It is important to note that the present and future perfect tenses are

marked using the same morphology. Their interpretation depends on the aspect of the lexeme; imperfect lexemes are interpreted as present tense (18), while perfect lexemes are interpreted as future tense (19). Since verbal aspect is inherent in the verb stem[2], it was considered part of the lexeme rather than coded as a separate variable.

(18) Present *gotować* 'to cook (imperf)'

1_{SG}	gotuj-ę	1_{PL}	gotuj-emy
2_{SG}	gotuj-esz	2_{PL}	gotuj-ecie
3_{SG}	gotuj	3_{PL}	gotuj-ą

(19) Future (perfect) *ugotować* 'to cook (perf)'

1_{SG}	ugotuj-ę	1_{PL}	ugotuj-emy
2_{SG}	ugotuj-esz	2_{PL}	ugotuj-ecie
3_{SG}	ugotuj	3_{PL}	ugotuj-ą

Although present and future perfect tense verbs are not marked for gender agreement, past tense verbs are (20). The past tense is also regular, meaning that all verbs are marked using the same morphemes.

(20) Past *rzucać* 'to throw'

	SG. MASC.	SG. FEM.	PL. MASC.	PL. FEM.
1	rzuca-ł-e-m	rzuca-ł-a-m	rzuca-l-i-śmy	rzuca-ł-y-śmy
2	rzuca-ł-e-ś	rzuca-ł-a-ś	rzuca-l-i-ście	rzuca-ł-y-ście
3	rzuca-ł	rzuca-ł-a	rzuca-l-i	rzuca-ł-y

Future imperfect tense verbs may or may not have gender agreement, depending on how they are constructed (21). The variation in future imperfect stems from the fact that it is made up of the future 'to be' which is marked for person and number, and is optionally followed by either

[2] Verbal aspect can be signalled in a number of ways: through stem vowel quality (compare *rzucać* (imperfect) 'throw continually' with *rzucić* (perfect) 'throw once'), through prepositional prefixes (compare *gotować* (imperfect) 'be cooking' with *ugotować* (perfect) 'cook'), or through a derivational morpheme (compare *dać* (perfect) 'give' with *dawać* (imperfect) 'continually give').

the infinite form of the verb (which does not mark for gender) or the past participle (which does mark for gender).

(21) Future (imperfect) *rzucać* 'to throw'

	NONE		MASC		FEM	
1SG	będę	rzuca-ć	będę	rzuca-ł	będę	rzuca-ł-a
2SG	będziesz	rzuca-ć	będziesz	rzuca-ł	będziesz	rzuca-ł-a
3SG	będzie	rzuca-ć	będzie	rzuca-ł	będzie	rzuca-ł-a
1PL	będziemy	rzuca-ć	będziemy	rzuca-l-i	będziemy	rzuca-ł-y
2PL	będziecie	rzuca-ć	będziecie	rzuca-l-i	będziecie	rzuca-ł-y
3PL	będą	rzuca-ć	będą	rzuca-l-i	będą	rzuca-ł-y

In sum, the four verbal morphology factors are: tense (past, present, future perfect, future imperfect); person (first, second, third); number (singular, plural); and gender (masculine, feminine, null). Each token was coded for these four factors based only on the morphology on the verb. For instance, if the subject was feminine (i.e. *ona* 'she') but the verb was not inflected for gender, the token was coded as "none" for grammatical gender. It is particularly important to accurately code for gender since not all tensed verbs make the distinction, and it has been argued in the past that greater subject agreement is linked to more null subject use.

Lexeme

Initially, each lexeme received a separate code based on its stem, including any derivational and aspectual markers. The lexemes were then regrouped in two ways and run separately in the analysis to determine their effect. The two types of lexical grouping were based on inflection class, and semantic content.

Grouping verbs by inflection class is motivated by McShane's (2009:115) observation that phonological ambiguity of person/number marking may lead to greater use of overt pronouns. Six main inflection classes (22) have been identified based on 1SG and 3SG present or future

perfect morphemes. Recall that present and future perfect tense use the same morphological system, and that past and future imperfect do not distinguish verbs based on inflection class.

(22)

	1SG	3SG
a.	*-ę*	*-e*
b.	*-ę*	*-(i)e*
c.	*-ę*	*-i*
d.	*-em*	*-e*
e.	*-am*	*-a*
f.	*-em*	*Ø*

Within these distinctions, (22a) would be the most phonetically ambiguous paradigm since denasalization at the end of a word causes the 3SG suffix *-e* /ɛ/ and 1SG suffix *-ę* /ɛ̃/ to sound identical in fast speech [ɛ]. Therefore, to support McShane's (2009) claim that phonetically ambiguous contexts lead to more overt subject use, we would expect fewer null subjects with verbs that inflect for the paradigm in (22a). Perhaps the least phonetically ambiguous paradigm is (22b), where the 3SG suffix *–e* causes palatalization on the preceding consonant. Therefore, in those verb the distinction between 1SG and 3SG lies not only in the quality of the final vowel but also in the quality of the stem consonant. Finally, it remains to be said that the paradigm in (22f) is represented by a single verb: *być* 'to be'.

Grouping verbs by lexical-semantic content, on the other hand, is motivated by work on Spanish null subjects (see for example Otheguy *et al.* 2007 and works cited therein), which found this to be a significant variable. In this case, lexemes are divided into three categories: mental/estimative verb, stative verb, and external action verb. Mental/estimative verbs were found to occur with the most null pronouns in Otheguy *et al.*'s study. There is perhaps a universal cognitive explanation for this; since mental verbs express thoughts and feelings, these

verbs may be targets of emphasis in situations where the speaker feels the need to stress that what follows is one person's opinion and not a fact.

Negation

Tokens were also coded for the presence of the negative morpheme *nie* immediately before the verb. The difference between negative and affirmative clauses has been found significant for second and third generation Heritage Russian speakers (Hollett 2010 in Nagy *et al.* 2010). One possible reason for this significance is syntactic; the verb is structurally separated from its subject by the negative morpheme.

Sentence type

Sentence type, or the distinction between interrogatives and statements, was also coded for. While there is no precedent for investigating this variable, the different syntax associated with interrogatives may affect choice of pronoun variant. For instance, interrogative morphemes, which immediately precede the subject, may strongly prefer to collocate with overt subject pronouns.

Subject continuity

For each token, the previous tensed verb was considered to determine whether the subject of that verb refers to the same or different entity. Subject continuity, also often referred to as switch reference or clause connection, has been found significant for all languages studied, including English. Its universal nature may be because it is a processing constraint; whenever speakers change topics, they want to do so as clearly as possible. In (23) below, the subject of

each verb (underlined) switches first from 'he' to 'we' and then back to 'he', yet the first subject switch does not trigger an overt pronoun.

(23) *Tak,* ***on*** *w środku* *był. On normalnie…* ***Ø*** *zwiedzalismy a* ***on*** *sobie spał.* ***Ø*** *Chodził po muzeum i sobie Ø znajdował miejsca do spania.*
"Yes, **he** was inside. He just… [***Ø*-we**] toured and **he** slept. [***Ø*-he**] walked around the museum and [***Ø*-he**] found places to sleep."
(0F16)

Initially, separate codes were assigned if the previous verb was uttered by the interviewer rather than the interviewee. Based on the pattern present in the data (see §5.1.1), these were later conflated so that the only distinction was whether the subject had the same or a different referent, regardless of who produced the previous verb. Note that in this study, the focus is on the referent rather than the form. Therefore, a referent was considered "same" in cases where the interviewee said "I" but the interviewer said "you" since in both cases the referent is the interviewee.

The coding of subject continuity was further complicated by the existence of what are arguably 'ergative' constructions in Polish. In these cases, the sentence does not have a nominative subject, but rather has a 'logical' subject expressed in the dative (c.f Gruszczyński & Bralczyk 2002p.189 'podmiot logiczny, podmiot w dopełniaczu'). Dative subjects often occur next to verbs denoting lack of something (24a), but similar constructions also exist with subjects in the genitive (24b), accusative (24c), and even instrumental (24d).

(24) a. *Nikogo nie ma w domu*
"Nobody's home." (lit. It hasn't anybody...)

b. *Ani zależy na ocenach*
"Anna cares about grades." (lit. It matters for Anna...)

c. *Mdli go po tych słodyczach*
"He's nauseous after that candy." (lit. It nauseates him…)

d. *Z tobą jest coraz gorzej*
"You're getting worse and worse." (lit. It's progressively worse with you.)

Initially these too were coded separately, so that if the previous verb had a logical subject that was coreferential with the subject of the token, the token received a unique code. Ultimately the pattern in the data showed that "same" logical subjects behaved like any other "same" subject (see §5.1.1). That is, while the grammatical subject of these constructions is technically "different", the logical subjects still act as primers for following sentences in the discourse.

Therefore, subject continuity ultimately had two distinctions in the final analysis: same and different. Tokens were coded as "same" if the subject of the token and the grammatical or logical subject of the previous verb uttered by the interviewee or interviewer were coreferential. Conversely, tokens were coded as "different" if the subject of the token and the grammatical subject of the previous verb uttered by the interviewee or interviewer were not coreferential.

Finally, it remains to be said that interjections (defined in §2.2.3) were not considered when coding for subject continuity. That is, if the previous tensed verb was part of an interjection, it was omitted and subject continuity was coded based on the verb before that.

Clause type

Finally, tokens were coded based on the type of clause they appeared in. Previous studies suggest that subordinate or coordinate clauses behave differently than main clauses (c.f. Otheguy *et al.* 2007; Morales 1997; Nagy *et al.* 2010). While these studies tend to distinguish clauses based on three broad categories (main, subordinate, coordinate), in this study each clause was coded based on the particular morpheme that heads it. The hope was that this coding method

would lead to previously undiscovered patterns and a better understanding of the exact role that clause type plays. For instance, individual clause morphemes can be indicative of a formal or informal register, signal the particular function of the clause, may be inherently semantically contrastive, or the form may simply be susceptible to collocations with overt subjects.

A list of each morpheme found in the data is given in (25).

(25) a. Main clause morphemes
Ø
to – optional signal of matrix clause in second position

b. Coordinate clause morphemes

i 'and'	*dlatego* 'that's why'
albo 'or'	*ale* 'but'
czy 'or'	*natomiast* 'but'
więc 'so'	*a* 'but'
tak że 'so'	

c. Subordinate clause morphemes

jak 'when'	*jak* 'how'
kiedy 'when'	*czy* 'if'
dopuki 'until'	*który* 'which'
ponieważ 'because'	*co* 'what'
dlatego że 'because'	*gdzie* 'where'
bo 'because'	*skąd* 'from where'
że 'that'	

Examining the patterns of these specific clause types may give insight into null subject use based on semantic or pragmatic function. Some morphemes above were later grouped based on their function and observed pattern (see §5.1.1).

4.1.2 Ethnic Orientation

As mentioned in §2.1.2, the discussion in all the sociolinguistic interviews centred on an ethnic orientation questionnaire made up of 37 questions. After collecting and coding the data

for all speakers, it became apparent that some revisions to the original questionnaire had to be made. One issue involved the types of answers speakers gave. Not only were some questions left unanswered, but some questions received the same answer from all speakers. Therefore, some questions were ultimately removed from the final ethnic orientation measures. Furthermore, the questions focused on speakers' attitudes only at the time of interview. However, life-experience is not static, and ethnic orientation can change drastically throughout a person's life. These considerations were the basis for the distinction between recent and childhood language environments. The following answer to the question "Are most of your friends Polish?" highlights this distinction particularly well:

> "Now, but not always. In the last five years since I moved back to Toronto. … I lived in the area of Kitchener-Waterloo for about 15 years. And in those times, I didn't have anybody, only family. My only Polish friends were my family. And when I came back to Toronto, I wanted, once again, to renew my culture." (2M47—my translation)

Ultimately, the original 37 questions were pared down to the following 15 measures of language contact (26):

(26) Ethnic orientation measures
Ethnic Identity
1. Speaker
2. Speaker's parents
3. Speaker's partner

Recent Linguistic Environment
4. Language used at work and/or school
5. Language used at home (with parents or partner)
6. Language used with friends

Childhood Linguistic Environment
7. Language used at school
8. Language used at home (with parents and/or siblings)
9. Language used with friends

Language Use

10. How well do you speak Polish?
11. How often do you speak Polish?
12. Which language do you prefer to speak?

Cultural awareness

13. Do you watch/listen/read Polish media?
14. Do you visit Poland? How often? For how long?
15. Are you aware of discrimination against Polish people?

A speaker's answer to each question was scored in one of three ways: 0 was given if the answer was oriented towards Canadian culture or the English language; 2 was given if the answer was oriented towards Polish language and culture; and 1 was given if the orientation was mixed. (See Appendix E for full coding schema.) The total number of points was then divided by the number of questions to get an average ethnic orientation score for each speaker. The result was that each speaker fit on an ethnic orientation continuum from 0 to 2, with 0 being completely English Canadian and 2 being completely Polish.

Two additional scoring methods were then devised: language mixing and linguistic confidence. The language mixing score attempted to test whether using both English and Polish in the same environment caused greater English influence than contact with English alone. It looked at 8 out of the 15 measures of language contact (27). The answers were then scored in one of two ways: 0 was given if both English and Polish were used; while 1 was given if either English of Polish were used exclusively. As with the ethnic orientation score, the total number of points was divided by the number of questions to get an average language mixing score. In this case, speakers fit on a language mixing continuum from 0 to 1, with 0 being someone who uses both English and Polish everywhere, and 1 being someone who keeps the two languages separate.

(27) Language mixing measures

Ethnic Identity

1. Speaker

Recent Linguistic Environment

2. Language used at work and/or school
3. Language used at home (with parents or partner)
4. Language used with friends

Childhood Linguistic Environment

5. Language used at school
6. Language used at home (with parents and/or siblings)
7. Language used with friends

Language Use

8. Which language do you prefer to speak?

Furthermore, a speaker's linguistic confidence was also measured. The linguistic confidence score examined the speaker's Polish language skills as well as the frequency of the speaker's use of Polish, and combined the researcher's assessment with the speaker's subjective assessment (28).

(28) Linguistic confidence measure

1) a. How well do you speak? (researcher's assessment)
   Good = 1
   Bad = 0
   b. Good enough? (speaker's assessment)
   Yes = +0
   No = +1
2) a. How often do you speak? (researcher's assessment)
   Almost every day = 1
   About once a week = 0
   b. How often do you speak? (speaker's assessment)
   Often = +0
   Rarely = +1

After dividing the total number of points by the number of questions, the result is an average linguistic confidence score from 0 to 2. This type of scoring creates three categories of speakers: speakers with a score of 0 do not speak well or often, but are not bothered by it;

speakers with a score of 1 speak relatively well and often, and they know it; speakers with a score of 2 speak well and often, but they feel they could do even better.

4.2 Hypotheses regarding individual factors within factor groups.

The literature outlined in §3.1 suggests a number of hypotheses about the significance of particular factors within the factor groups examined in the current study. As has already been mentioned in §3.3, much of the previous research on null subject languages has pointed to the significance of person and/or number agreement. However, the particular type of agreement that is likely to appear with null subjects depends on the language. For instance, Russian was found to exhibit more null pronouns in the 3rd person, while Cantonese found 1st person subjects more likely to be null. Italian, on the other hand, exhibited significance depending on number agreement, with plural subjects being most often null. Similarly, studies on Spanish found more null with 1st and 3rd person plural subjects than with any singular subjects. Because of this varying behaviour across null subject languages, it is not possible to hypothesize about these factors with respect to Polish. However, it is interesting to note that in her study of English, Harvie (1998) found most null subjects occur in the 1st person. Therefore, if English has had influence on the immigrant generations, it may be that 1st person subjects will most often be null for those speakers.

Second, if it is true that richer agreement morphology allows for more null subjects, the prediction would be that verb forms which are more marked with respect to subject agreement will exhibit more null subject than those that are not. Applied to Polish, this would predict that tense/aspect combinations that inflect for gender (i.e. all past tense forms, and some future imperfect forms) would have more null subjects than verb forms that are not (i.e. all present

tense forms, and all future perfect forms). This hypothesis has already been borne out based on the categorically null behaviour of conditional tense/aspect forms, which do inflect for gender agreement[3]. Furthermore, while verb forms that do not inflect for gender are expected to have fewer null subjects overall, it is also predicted that verb forms inflected for masculine gender (which are morphologically less marked) will have fewer null subjects than verb forms inflected for feminine gender (which is typically marked with the vocalic suffix *–a*).

Further hypotheses based on inflection involve phonetic ambiguity of particular lexical inflection classes. If it is true that phonetically ambiguous forms are more likely to exhibit overt pronouns, we would expect verbs belonging to the (*ę ~ e*) inflection class to have the least null subject out of all the inflection classes. Conversely, all other inflection classes are expected to behave equally with respect to null subjects since they are all unambiguous.

Another lexically-based factor, semantic content, is also expected to be significant. Based on Otheguy *et al.* (2007), action verbs are expected to occur with the most null subjects, while mental/estimative verbs are expected to occur with the least null subjects. Mental/estimative verbs are also more susceptible to pragmatic emphasis and contrast. Similarly, negative constructions and interrogative constructions may also be susceptible to emphasis, and as such may also appear with fewer null subjects.

Within subject continuity, it is expected that different subjects will have fewer null realizations than same subjects. However, recall that, at step 2 of Heine & Kuteva's (2005) process of grammaticalization, extension of overt subjects to new (non pragmatic) contexts suggests a reduction in the significance of the subject continuity variable. Recall also Johanson's

[3] Gender agreement on conditional verbs is marked in the exact same way as on past tense verbs by using identical morphology. It is important to note that while the categorical behaviour of conditional forms supports the hypothesis that richer agreement is linked to null subject use, the fact that past tense forms are not categorical is already problematic, especially if both form types use the same gender morphemes.

(2008) claim that in grammatical replication the replica language moves towards, but does not fully become, the model language. Therefore, although subject continuity has also been found to be the most significant factor in English, I would expect subject continuity to become less significant in subsequent generations as a result of English influence.

Although clause type alone has been found to be significant for some languages (relative subordinate clauses have more overt subject in Spanish according to Morales 1997; the second conjuncts in conjoined clauses have more null subject or fewer null subject in Russian and Cantonese respectively according to Nagy *et al.* 2010), it is the combination of clause type and subject continuity that is expected to exhibit a significant pattern. For instance, Nagy *et al.*'s (2010) results for null subjects in English showed that the difference between same and different subjects is greater for conjoined clauses than for main clauses. Furthermore, Lindseth claims that in null subject languages such as Polish only null pronouns can function as bound variables (29).

(29) a. Oni$_1$ myślą że oni$_2$ są mądrzy.
they think that they are smart (1 ≠ 2)

b. Oni$_1$ myślą że Ø$_2$ są mądrzy.
they think that they are smart (1 = 2)

Because a null subject in a subordinate clause may be interpreted as coreferential with the subject of the matrix clause, it is expected that different referents in subordinate clauses will exhibit the fewest null subjects. At the same time, it is also expected that a pattern will emerge with respect to particular clause morphemes, with semantically contrastive morphemes co-occurring with the least null subject pronouns. This is based on the assumption that overt pronouns are contrastive.

Table 4.1

Summary of hypotheses regarding significant linguistic variables

Factor group	Factor expected to have the most null pronouns	Factor expected to have the least null pronouns
Person+number	• 1st person subjects • plural subjects	
Tense+gender	• past tense • feminine gender	• present tense & future perfect
Inflection class	• *ę~e* phonetically ambiguous paradigm	
Semantic content		• mental/estimative verbs
Subject continuity	• same as previous clause	• different from previous clause
Clause morpheme type	• same subjects in conjoined clauses	• different subjects in subordinate clauses • semantically contrastive morphemes

Finally, there are some discrepancies in the findings on the effects of speakers' contact with and attitude toward English on the null subject rate. Otheguy *et al.* (2007) show that increased contact with English results in fewer null subjects while Bayley & Pease-Alvarez (1997) suggest that greater depth of ties to the majority language and culture results in more null subjects. Conversely, Lustanski's (2009a) findings indicate that ethnic orientation will not have any effect on the degree of standard language maintenance. Therefore, while in general I expect first generation speakers to be less affected by English than second generation speakers, I believe that certain social factors will be more significant than this generalization would suggest. Based on Niedzielski (1997), two social factors that influence language maintenance in particular are the amount of language education received by immigrants, and the amount of external pressure they have to assimilate. He offers some anecdotes from his experience to clarify these factors. When speaking to Poles living in New Zealand, who he found to have the greatest language maintenance out of all the immigrant communities he interviewed, he was continually asked about the "proper" way to say something. He also found that Poles in the U.S., who were

permitted to speak Polish at their workplace, exhibited less language maintenance than Poles residing in France, who were pressured to use only French at work. Therefore, I expect that linguistic confidence and the desire to speak "proper" Polish will resist English influence. In other words, speakers who wish to improve their use of Polish will have a higher null subject rate than those who do not. On the other hand, opportunities to use the Polish language in addition to English in a variety of social contexts will attract English influence. This means that speakers who use both English and Polish in the same social contexts will use fewer null subjects.

5.0 RESULTS

This section documents the analyses performed on the data, outlining the statistical information and the patterns that these statistical models suggest. I begin by looking at the distribution of tokens in the language data with respect to a number of social and linguistic factors (§5.1). I then turn to a multivariate analysis to determine which of these social and linguistic factors are significant in the grammar (§5.2). Subsequently, I separate the data into three groups based on the speakers' immigrant generation, and perform multivariate analyses aimed at determining the difference between the three groups (§5.3). Finally, I analyze the ethnic orientation data and how it correlates to each speaker's null subject rate (§5.4).

5.1 Distributional analysis

5.1.1 Overall distribution of the corpus

The following table (5.1) shows an overall distribution of null and overt pronouns in the entire corpus. This includes all tokens that fit in the envelope of variation (as defined in §2.2), and is made up of data from all three generations of speakers.

Table 5.1

Overall distribution of null and overt pronouns			
Null		Overt	
%	N	%	N
76.5	732	23.5	225
Total N	957		

The distribution of these 957 tokens shows that, on average, null pronouns are three times more likely to be used than overt pronouns by Polish speakers. In contrast, studies of null

pronouns in English (e.g. Harvie 1998) report that a null pronoun rate for English cannot be established because they are so rare. In the following section, I discuss the use of null subject pronouns based on linguistic factors (as explained in §4.1) as well as social factors.

5.1.2 Social factors

Each token in the data was identified by speaker. As mentioned in §2.2.1, speakers were grouped based on the immigrant generation they belong to.

Table 5.2

Distribution of Null subject variants by generation since immigration		
Generation	%	N
Homeland speakers	80.7	197
First generation	77.4	385
Second generation	73.3	375
Total N	957	

Table 5.2 above shows that the percentage of null subject pronoun variants falls with each subsequent generation. The homeland speakers, who have had no contact with English, have the greatest null subject rate; while the second generation speakers, who have arguably had the greatest contact with English, have the lowest null subject rate.

Another social variable is speaker sex.

Table 5.3

Distribution of Null subject variants by speaker sex		
Sex	%	N
Female	77.6	361
Male	75.8	596
Total N	957	

Table 5.3 above shows that, overall, there does not appear to be any effect with respect to speaker sex. However, when this factor is cross-tabulated with immigrant generation (Table 5.4), a different pattern emerges.

Table 5.4

Cross-comparison by speaker generation of Null variants by speaker sex

	Homeland		First		Second	
Sex	%	N	%	N	%	N
Female	81	197	85	54	68	110
Male	--	--	76	331	75	265
Total N	197		385		375	

While the null subject rate for males appears to be steady across first and second generation speakers, there is a large difference (17%) between first generation and second generation females. On the other hand, there appears to be little difference between homeland females and first generation females. This distribution suggests that only females in the second generation are converging with English. However, without data from additional speakers and further statistical testing, we cannot tell whether these sex and generation differences are significant.

5.1.3 Person and number

Grammatical person and/or number have been found to be significant in null subject languages previously studied. The distribution data for Polish shows that while there is a relatively large difference in rate among different grammatical persons with third person tokens having the fewest null subjects (Table 5.5), the difference in rate between the two grammatical numbers is quite small (Table 5.6).

Table 5.5

Distribution of Null subject variants by grammatical person		
Person	%	N
Second	81.0	63
First	79.6	661
Third	66.5	233
Total N	957	

Table 5.6

Distribution of Null subject variants by grammatical number		
Number	%	N
Plural	78.0	223
Singular	76.0	734
Total N	957	

When these two factors are cross-tabulated (Table 5.7), it suggests that they may not be treated separately. Focusing on the first-person and third-person tokens, we find that in each case there are more null subjects in the plural than in the singular.

Table 5.7

Cross-comparison by grammatical person of Null variants by grammatical number						
	First		Second		Third	
Number	%	N	%	N	%	N
Singular	79	536	84	57	63	141
Plural	84	125	50	6	72	92
Total N	661		63		233	

However, due to the nature of the interviews, there are very few second person tokens, and in particular very few second person plural tokens. Therefore, these six tokens are ultimately eliminated from future multivariate analyses, and the grammatical person and number factors are combined to form one factor group (see §5.2). A factor group that intersects

grammatical person and number is also useful in revealing specific patterns and differences between different generations (see §5.3).

5.1.4 Gender and Tense

Recall from §4.1.1 that grammatical gender agreement is not present on all verb forms. Recall also that null subjects are commonly believed to correlate with greater subject agreement on the verb. Therefore, it would follow that those verb forms that have gender agreement in addition to person and number agreement would be more likely to occur with null subjects than those that do not.

However, as the distribution in Table 5.8 shows, the opposite is true; verb forms that do not have gender agreement exhibit a greater null subject rate. Furthermore, masculine gender agreement (which is marked by a null morpheme) has a greater null subject rate than feminine agreement (which is marked by a vocalic morpheme).

Table 5.8

Distribution of Null subject variants by grammatical gender

Gender	%	N
None	79.2	501
Masculine	77.4	319
Feminine	64.2	137
Total N	957	

At this point, it is important to determine whether this distribution pattern is true for all speakers. Since only female speakers are able to produce first person feminine tokens, we must ensure that these types of tokens are not driving down the rate. Table 5.9 below shows that both male and female speakers have the same type of pattern with respect to grammatical gender. In

fact, male speakers are more likely than female speakers to use null subjects with feminine agreement.

Table 5.9

Cross-comparison by gender agreement of Null variants by speaker sex						
	None		Masculine		Feminine	
Speaker sex	%	N	%	N	%	N
Female	83	153	83	86	67	122
Male	78	348	76	233	40	15
Total N	501		319		137	

Lack of speaker effect is further apparent if we cross-tabulate grammatical gender with person+number agreement. If the higher rate of overt subject pronouns with feminine agreement were in any way dependent on the speaker, then we would expect significantly fewer null subjects in first person singular contexts. However, masculine (77%) and feminine (71%) rates in this context are similar. Therefore, the null subject rate in feminine contexts is not a reflex of women overusing overt pronouns when referring to themselves. In third person singular contexts, on the other hand, the null subject rates of masculine (72%) and feminine (44%) agreement differ dramatically; a speaker is more likely to overtly say *ona* 'she' than *on* 'he'.

Table 5.10

Cross-comparison by gender agreement of Null variants by person+number						
	None		Masculine		Feminine	
Person+number	%	N	%	N	%	N
1 sg.	82	281	77	171	71	84
2 sg.	95	43	67	3	45	11
3 sg.	64	73	72	43	44	25
1 pl.	84	45	87	69	64	11
2 pl.	100	1	0	2	67	3
3 pl.	69	58	74	31	100	3
Total N	501		319		137	

Finally, recall that that the presence of gender agreement depends on verb tense. Table 5.11 below shows that the past tense, which always has gender agreement, has the fewest null subjects. Future imperfect, which sometimes allows for gender agreement, has fewer null subjects than future perfect, which does not allow for gender agreement.

Table 5.11

Distribution of Null subject variants by tense		
Tense	%	N
Future perfect	90.0	30
Future imperfect	84.6	13
Present	78.3	466
Past	73.4	448
Total N	957	

A cross-tabulation of gender and tense factors (Table 5.12) further reveals that the only overt subject tokens in the future imperfect tense are ones with gender agreement (i.e. "Masculine" or "Feminine"). All future imperfect tokens without gender agreement ("None") have null subjects. Overall, since forms with greater subject agreement exhibit *fewer* null subjects, this disproves claims that greater agreement on the verb is correlated to greater null subject use.

Table 5.12

Cross-comparison by gender agreement of Null variants by tense						
	None		Masculine		Feminine	
Tense	%	N	%	N	%	N
Present	78	466	--	--	--	--
Past	--	--	77	315	64	133
Future perfect	90	30	--	--	--	--
Future imperfect	100	5	75	4	75	4
Total N	501		319		137	

The lack of balance with respect to tense and gender marking inherent in Polish grammar results in a number of empty cells. Therefore just as with grammatical person and number, the tense and gender factors will be combined to form one factor group in subsequent multivariate analyses. Since both future tenses have the fewest tokens, and since they both have very high null subject rates, they will be conflated into one "future" factor. Furthermore, present tokens will remain categorized as "present", while past tokens will be divided into "past masculine" and "past feminine" categories.

5.1.5 Negation and Sentence type

Two minor factors in the data involve the presence of negation in the clause, and the syntactic sentence type. Table 5.13 shows that negated verbs have a slightly lower rate of null subject pronouns.

Table 5.13

Distribution of Null subject variants by presence of negation

Negation	%	N
Affirmative	76.9	848
Negative	73.4	109
Total N	957	

However, a greater difference in null subject rates is present in the distribution by sentence type (Table 5.14). Questions are much less likely than statements to have null subjects, although this may be due only to the small number of tokens in the cell. Nonetheless, in order to ensure that the data was balanced, the 17 question tokens were removed from the data and this factor was removed from further analysis.

Table 5.14

Distribution of Null subject variants by sentence type		
Sentence	%	N
Statement	76.8	940
Question	58.8	17
Total N	957	

5.1.6 Inflection class

Another common belief present in the literature is that surface ambiguity of the verb form may be correlated to lower null subject rates. Table 5.15 below shows the distribution of null subjects by verb inflection class, one of which (*ę ~ e* in bold) may be phonetically ambiguous and as such is expected to exhibit fewer null subjects than other verb inflection classes.

Table 5.15

Distribution of Null subject variants by verb inflection class		
Inflection	%	N
ę ~ $^{(i)}$e	85.4	123
em ~ e	77.3	44
am ~ a	76.2	307
ę ~ i	76.0	246
ę ~ e	**74.4**	**125**
irregular (BE)	70.5	112
Total N	957	

Overall, the irregular inflection class (BE) has the fewest null subjects, which is unexpected considering it is one of the least ambiguous paradigms. However, this may be due to the frequency with which this particular verb is used. The other less ambiguous class (*ę ~ $^{(i)}$e*), where third person singular inflection causes palatalization on the preceding consonant, is also the class with the most null subjects. This is expected, since it is easier to correctly infer the subject of a verb whose agreement morphology is unambiguous. All other inflectional classes

are relatively similar in terms of subject pronoun behaviour. However, note that, next to the irregular paradigm, the most ambiguous class (*ę ~ e*) does in fact have the fewest null subjects. This suggests that ambiguity does lead to overt subject use.

However, further analysis indicates otherwise. Recall that verbs are distinguished by inflectional paradigm only in the present and future perfect tenses. Therefore, it is important to make sure that the distribution in Table 5.15 above is not skewed by tokens in other tenses. Table 5.16 below shows a cross-tabulation of tense and inflection class factors.

Table 5.16

Cross-comparison by tense of Null variants by verb inflection class

	Present		Future		Past Masc.		Past Fem.	
Inflection	%	N	%	N	%	N	%	N
ę ~ $^{(i)}$e	86	42	100	13	84	50	78	18
em ~ e	79	21	--	--	67	9	86	7
ę ~ e	**79**	**21**	100	3	78	32	50	20
am ~ a	78	175	79	19	77	81	62	32
ę ~ i	76	106	100	6	78	91	67	43
irregular (BE)	76	45	50	2	73	52	46	13
Total N	466		43		315		133	

The small number of tokens in the future tense makes it difficult to draw any conclusions based on that subset of data. However, it is still possible to examine the patterns in present tense data. There, the most ambiguous paradigm (*ę ~ e*) is not any less likely to have null subjects than any other paradigm. Nonetheless, the least ambiguous class (*ę ~ $^{(i)}$e*) is still found to be the class with the most null subjects. While all other inflection classes fall within a small range (76-79%), (*ę ~ $^{(i)}$e*) has a rate of 86%. Therefore, while the presence of morphological ambiguity *per se* does not drive down the null subject rate, strong lack of ambiguity does appear to drive up the null subject rate.

5.1.7 Semantic content

Lexemes can further be divided based on their semantic content: namely, whether they denote an external action, a physical state, or a mental state. Table 5.17 below shows a progression from action verbs (which have the highest null subject rate), to stative verbs, to mental verbs (which have the lowest null subject rate).

Table 5.17

Distribution of Null subject variants by lexical content of verb		
Content	%	N
Action	79.7	477
Stative	75.9	319
Mental	68.3	161
Total N	957	

5.1.8 Subject continuity and Clause type

As discussed in §4.1.1, originally tokens were coded based on a five-way distinction. Table 5.18 below shows the distribution of null subjects in these five categories.

Table 5.18

Distribution of Null subject variants by subject continuity		
Continuity	%	N
object of ergative **(S)**	94.1	17
same (interviewee) **(S)**	84.5	452
same (interviewer) **(S)**	80.2	86
different (interviewee) **(D)**	66.4	378
different (interviewer) **(D)**	58.3	24
Total N	957	

The distribution above shows that a preceding object of an ergative subject-less clause, which is coreferential with the subject of the token verb, appears to have a strong priming effect. The number of tokens in this category is quite small, since these types of constructions are

limited to particular verbs and expressions, and it is therefore rare for regular clauses to be preceded by them. Nonetheless, the verb following these ergative constructions are most likely to appear with a null subject.

Furthermore, if we compare "interviewee" and "interviewer" tokens with either a same or different referent we find that, in conversational turn-taking, a speaker is less likely to use a null subject at the beginning of their turn than at a later stage. This difference appears to be stronger for "different" referents than for "same" referents. Nonetheless, even at the beginning of a speaker's turn, a same referent is much more likely to have null subjects than a different referent. Therefore, for all intents and purposes, there is a two way distinction in subject continuity: (S) "same" versus (D) "different" referent. "Same" tokens include those preceded by coreferent objects of ergative, and coreferent subjects spoken by either the interviewee or interviewer; "different" tokens include those preceded by non-coreferential subjects spoken by either the interviewee or interviewer. This two-way distinction will be used in the multivariate analyses.

With respect to clause type, recall that each clause was coded based on the morpheme that heads it. However, most analyses of null subject languages make a three way clause type distinction: main, subordinate and coordinate. If the data is recoded into these three categories, the distribution in Table 5.19 emerges.

Table 5.19

Distribution of Null subject variants by clause type

Clause type	%	N
Main	77.6	468
Subordinate	77.6	254
Coordinate	73.2	235
Total N	957	

The distribution above shows little, if any, difference between the three clause types. However, it is possible that these clause distinctions may pattern differently with respect to subject continuity. This possibility has been suggested by Nagy *et al.*'s (2010) findings for English null subjects.

Table 5.20

Cross-comparison by subject continuity of the distribution of Null variants by clause type

	Same		Different	
Clause type	%	N	%	N
Main	85	284	67	184
Subordinate	85	142	68	112
Coordinate	82	130	63	105
Total N	401		556	

The cross-tabulation of clause type and subject continuity above (Table 5.20) shows that all three clause types behave similarly with respect to the two-way distinction in subject continuity.

However, differences in null subject rate emerge when each clause morpheme is examined individually (Table 5.21). Morphemes are glossed and identified by clause type: (M) main/matrix clause; (C) coordinate clause; (E) embedded clause.

Table 5.21

Distribution of Null subject variants by clause morpheme		
Clause morpheme	%	N
kiedy, dopuki – 'when, until' **(E)**	100.0	8
ponieważ, dlatego że – 'because' **(E)**	100.0	7
jak, czy – 'how, if' **(E)**	85.7	14
jak – 'as, when' **(E)**	84.7	72
że – 'that' **(E)**	83.3	60
i – 'and' **(C)**	78.7	108
to – matrix clause **(M)**	78.4	102
Ø – main clause **(M)**	77.3	366
albo, czy – 'or' **(C)**	75.0	8
więc, tak że, dlatego – 'so, that's why' **(C)**	73.3	30
ale, natomiast – 'but' **(C)**	69.8	63
Relative clause **(E)**	67.9	28
bo – 'because' **(E)**	61.5	65
a – 'but' **(C)**	57.7	26
Total N	957	

Matrix clauses preceded by *to* and main clauses that appear on their own have very similar rates. It appears that the position of main/matrix clause does not affect null subject rates.

Most types of embedded clauses have a higher null rate than matrix clauses. However, two particular contexts—clauses following *bo* 'because' and relative clauses—have a lower null rate. Furthermore, the two categorically null environments, despite having low token numbers, are interesting when compared to their semantic counterparts in the data. Namely, the more formal *kiedy* 'when' occurs with more null subjects than the less formal *jak* 'when'; while the more colloquial *bo* 'because' occurs with significantly fewer null subjects than the formal *ponieważ* and *dlatego że* 'because'. Recall McShane's (2009:107) summary that there are considerably more overt subjects in spoken Polish than in the written language. Together, these results suggest that formality plays a role in null subject use, with more formal language employing more null subjects than more informal language.

Finally, coordinate conjoined phrases meaning 'and' or 'or' have a similar rate to main clauses, whereas more contrastive conjoined phrases meaning 'but' have a lower null pronoun rate. This may be a reflex of the inherent contrastive property of 'but'; that is, perhaps these contexts have more different subject referents. A cross-tabulation of clause morpheme and subject continuity (Table 5.22) shows that this is not necessarily true.

Table 5.22

Cross-comparison by subject continuity of the distribution of Null variants by clause morpheme

	Same		Different	
Clause morpheme	%	N	%	N
jak, czy – 'how, if'	80	10	100	4
jak (incl. *kiedy, dopuki*) – 'when'	88	40	85	40
że – 'that'	92	39	67	21
to – matrix clause	88	59	65	43
Ø – main clause	84	225	67	141
i – 'and'	83	64	73	44
albo, czy – 'or'	80	5	67	3
więc, tak że, dlatego – 'so, that's why'	78	18	67	12
Relative clause	69	16	67	12
ale, natomiast – 'but'	80	35	57	28
bo (incl. *ponieważ, dlatego że*) – 'because'	84	37	46	35
a – 'but'	88	8	44	18
Total N	401		556	

In "same" contexts, most clauses hover between an 80% and a 90% null subject rate. The only major deviation occurs in relative clauses, which are more likely to have overt subjects than all other clauses. In "different" contexts, most clauses hover at a 67% null subject rate. However, unlike "same" contexts, "different" contexts exhibit more variety among clause morphemes. The major deviations from the median null subject rate lie in contrastive conjunctions meaning 'but', and in subordinate clauses meaning 'because'. Both types of clause morphemes exhibit fewer null pronouns. On the other hand, conditional-type subordinate

clauses (i.e. "how, if, when") exhibit a higher null subject rate. Further, note that while *a* and *ale* 'but' pattern in similar ways, there is a greater difference between "same" and "different" contexts for *a* clauses.

An additional pattern emerges as a result of these discrepancies in "same" versus "different" contexts. Namely, relative clauses and conditional subordinate clauses do not exhibit major differences in rate depending on subject continuity. Null subject pronouns are just as likely to occur in different contexts as in same contexts for clauses headed by these morphemes. It is interesting that all these clause types form a class, in that they are all headed by question morphemes.

Based on the patterns above, certain clause morphemes were conflated into one factor, provided that there were few tokens of the particular morpheme and it could be said to pattern with another morpheme. Clause morphemes with a large number of tokens remained separate to leave open the possibility of examining their particular patterns across generations. The final clause morpheme distinctions were: (1) *jak, czy, kiedy, dopuki* 'how, if, when'; (2) *że* 'that'; (3) *to* matrix clause; (4) Ø main clause; (5) *i, albo, czy, więc, tak że, dlatego* 'and, or, so'; (6) relative clause; (7) *ale, natomiast, a* 'but'; (8) *bo, ponieważ, dlatego że* 'because'.

5.2 Multivariate analysis

As discussed in the previous section, tokens with second person plural agreement and from interrogative sentence types were excluded from multivariate analyses. The overall distribution of the remaining tokens is given in Table 5.23. Note that the distribution is not significantly different from the distribution for all tokens in the corpus given in §5.1.1.

Table 5.23

Overall distribution of null and overt pronouns, after exclusions

Null		Overt	
%	N	%	N
77	719	23	215
Total N	934		

Multivariate analyses were conducted using GoldVarbX to determine the significance of linguistic and social factors to the probability of null subject pronouns. The program identifies which previously defined factor groups are statistically significant ($p<0.05$) in predicting the dependent variable, and lists these groups from most predictive to least predictive. The program also calculates the corrected mean and the factor weights for each factor in a factor group. The corrected mean indicates the average rate of the dependent variant, all else being equal. The factor weights indicate whether a particular factor favours (>0.5) or disfavours (<0.5) the variant; in the present case, the values indicate whether the factor tends to drive up or drive down the average null subject rate. The numeric range between the highest and lowest factor weight within one factor group is an additional indication of the factor group's relative strength.

In the present study, nine factor groups were included in the final multivariate run: (1) subject continuity; (2) person+number; (3) immigrant generation; (4) tense+gender; (5) semantic content of lexeme; (6) clause morpheme; (7) verbal inflection class; (8) speaker sex; and (9) negation. The run identified six factor groups as significant. From most significant to least significant these factors are: subject continuity; person+number; immigrant generation; tense+gender; semantic content of lexeme; and clause morpheme. Table 5.24 summarizes the factor weights of each factor within these factor groups.

Table 5.24

Multivariate analyses of the contribution of internal and social factors selected as *significant* to the probability of null subject variants in Polish speakers

Corrected mean:			0.807
Total N			934
	Factor weight	%	N
Subject continuity			
Same	.62	84	546
Different	.34	67	388
Range	*28*		
Person+Number			
2 sg.	.65	88	49
1 pl.	.64	85	123
1 sg.	.55	79	531
3 pl.	.38	72	92
3 sg.	.24	63	139
Range	*41*		
Generation			
Homeland	.65	81	193
First	.52	78	371
Second	.40	74	370
Range	*25*		
Tense			
Future	.68	91	42
Present	.56	79	450
Past masculine	.47	78	312
Past feminine	.32	64	130
Range	*36*		
Semantic content			
Action	.56	81	467
Stative	.48	76	309
Mental	.36	68	158
Range	*20*		
Clause morpheme types			
'if/when'	.67	87	93
'that'	.60	83	60
'and'	.53	78	145
Main clause	.49	78	354
Matrix clause with *to*	.48	78	100
'but'	.40	67	84
'because'	.40	66	70
Relative clause	.31	68	28
Range	*36*		

Out of the nine factor groups defined for the final multivariate run, three were found to be non-significant: verbal inflection class, speaker sex, and negation. The factor weights for these non-significant factor groups are summarized in Table 5.25. The factor weights reported in these two tables further support the patterns observed in the distribution data, discussed in §5.1.

Table 5.25

Multivariate analyses of the contribution of internal and external factors selected as *not significant* to the probability of null subject variants in Polish speakers

Corrected mean:			0.810
Total N			934
	Factor weight	%	N
Inflection class			
ę ~ $^{(i)}$e	.60	87	120
em ~ e	.58	76	41
ę ~ i	.51	77	240
am ~ a	.49	77	298
ę ~ e	.48	74	125
Irregular (BE)	.42	70	110
Speaker sex			
Female	.57	78	355
Male	.46	77	579
Negation			
Positive	.50	77	826
Negative	.48	74	108

The factor groups and factor specifications that comprised the multivariate analysis presented in this section was chosen as the best fit for the current data (log likelihood=-439.190). Other configurations were also tested. For instance, the current multivariate analysis included a factor group that intersects person and number agreement. An analysis where this factor group was replaced by two factor groups that separate person and number was also performed. The results of this analysis indicate that both person agreement and number agreement are significant. However, person is more significant (2nd out of 7 factors) than number (6th out of 7 factors). The

log likelihood for this analysis (-439.399) when compared to the log likelihood of the final analysis presented above (-439.190) indicates that intersecting person and number agreement provides a slightly better fit.

Further analyses were also done to test the significance of clause type. Multivariate analyses that replaced the above clause morpheme factor specifications with a general three-way (main, conjoined, subordinate) clause type factor group did not select this group as significant. This type of analysis was also a worse fit for the data (log likelihood=-446.373), indicating that specific clause morpheme distinctions are more predictive than general clause type distinctions with respect to null subject pronouns.

Finally, the current analysis shows that inflection class is not significant. The results up until now suggested that while presence of ambiguity (*ę~e*) does not cause decreased use of null subjects, strong absence of ambiguity (*ę~$^{(i)}$e*) may cause increased use of null subjects. Nonetheless, multivariate analyses that regrouped the factors into fewer categories (e.g. *ę~$^{(i)}$e* vs. everything else; or (*ę~$^{(i)}$e*) vs. BE vs. everything else) did not find the factor group significant. Lack of statistical significance for any configuration and grouping of inflection class factors proves that the presence or absence of phonetic ambiguity on the verb form does not affect the use of null subject pronouns.

5.3 Comparison of results by speaker generation since immigration

Tokens from each of the three speaker generations (homeland, first, second) were isolated and subjected to separate multivariate analyses to determine whether there are differences in the grammars of these groups of speakers. Overall, the distribution of tokens (Table 5.26) shows that homeland speakers have a much higher null subject rate than both immigrant generations,

and that second generation immigrant speakers have a lower null subject rate than first generation immigrant speakers.

Table 5.26

Distribution of Null subject variants by generation since immigration		
Generation	%	N
Homeland speakers	80.8	193
First generation	78.4	371
Second generation	73.5	370
Total N	934	

The multivariate analyses show that a different configuration of factor groups is found significant for each generation. Table 5.27 lists the significant factor groups in each generation from most significant to least significant.

Table 5.27: Factor groups selected as significant for each immigrant generation

HOMELAND	FIRST GENERATION	SECOND GENERATION
1. Subject continuity	1. Subject continuity	1. Semantic content
2. Person+Number	2. Person+Number	2. Inflection class
3. Tense	3. Inflection class	3. Tense
		4. Person+Number
		5. Subject continuity

Based on the significant factors chosen, homeland and first generation speakers are very similar. Both have subject continuity and person+number as the two most significant factors. The only difference between those two speaker groups is that inflection class replaces tense as the third most significant factor for first generation speakers. On the other hand, second generation speakers exhibit a very different pattern. Most interestingly, these speakers have subject continuity and person+number as the two least significant factors out of five. Inflection class and tense are also significant, just as in first generation and homeland speakers respectively. However, the most significant factor for second generation speakers is semantic

content—a factor that is not significant at all for homeland and first generation speakers. Overall, these patterns show that speakers who were born and raised in Poland (i.e. homeland and first generation) are more similar to each other than speakers who have experienced prolonged contact with English (i.e. first generation and second generation).

Table 5.28 below shows a comparison by generation of significant factors[4]. Factors not found significant for a particular generation appear in square brackets. Distribution data by generation can be found in Appendix B. While for the most part the factors in each factor group are ranked similarly across the generations, some differences suggest that grammatical reanalysis is taking place.

The most obvious difference among the three generations is in their treatment of subject continuity. The range for subject continuity becomes smaller with each subsequent generation since immigration, showing that overall this particular factor is losing significance with more English contact.

[4] Recall from §5.2 that a factor group that intersects person and number agreement was found to be a slightly better fit for the data. This is also true for analyses looking at each individual generation. However, it is important to note that while analyses performed on all data found both person and number agreement individually significant, this was not true of analyses performed on data from individual generations. Whenever person and number agreement were run as separate factor groups, only person was selected as significant.

Table 5.28

Cross-generational comparison of multivariate analyses of the contribution of internal and external factors selected as *significant* to the probability of Null subjects

	Homeland		First		Second	
Corrected mean:	0.879		0.820		0.772	
Total N	193		371		370	
	Factor weight	N	Factor weight	N	Factor weight	N
Subject continuity						
Same	.73	101	.66	215	.56	230
Different	.26	92	.29	156	.41	140
Range	*47*		*37*		*15*	
Person+Number						
2 sg.	.44	13	.68	23	[1]	13
1 pl.	.77	38	.66	47	.54	123
1 sg.	.66	67	.52	191	.53	273
3 pl.	.26	20	.41	44	.42	28
3 sg.	.23	55	.33	66	.16	18
Range	*54*		*35*		*38*	
Tense						
Future	.79	17	[.55]	16	[1]	9
Present	.58	74	[.50]	185	.60	191
Past masculine	.58	45	[.55]	148	.43	119
Past feminine	.26	57	[.20]	22	.30	51
Range	*55*				*30*	
Inflection class						
em ~ e[5]	[.81]	7	.22	13	.78	21
ę ~ ^{i}e	[.73]	36	.47	43	.78	41
ę ~ i	[.55]	56	.50	104	.45	80
ę ~ e	[.49]	25	.34	42	.58	58
Irregular (BE)	[.33]	27	.37	31	.39	52
am ~ a	[.30]	42	.62	138	.38	118
Range			*40*		*40*	
Semantic content						
Action	[.50]	104	[.51]	196	.60	167
Stative	[.58]	61	[.49]	115	.54	133
Mental	[.33]	28	[.48]	60	.23	70
Range					*37*	

[5] Note that the large discrepancies in Factor Weight values between generations may be due to the small number of tokens in this category.

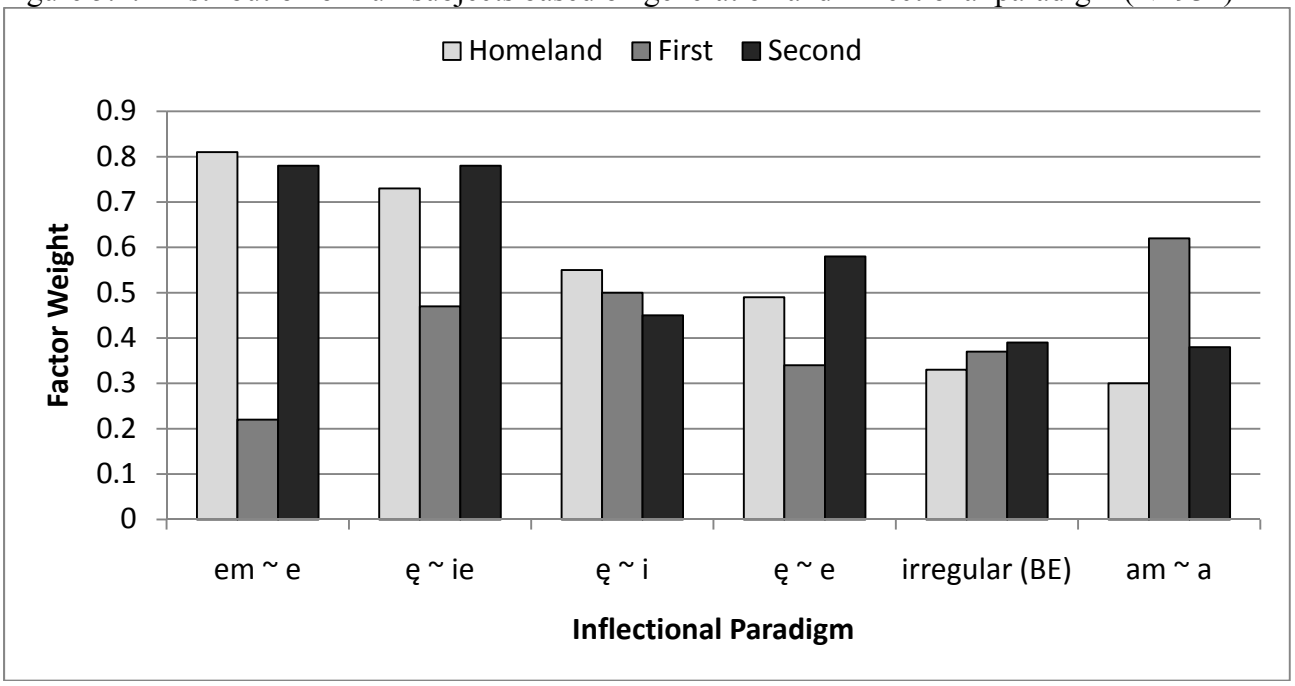
The situation with person+number is slightly more complicated. For one, the null rate in second person singular tokens actually rises to become categorical for second generation speakers. This may simply be due to the distribution in the data: there are generally few second person singular tokens; and perhaps second generation speakers always used it with a generic reference (which is more commonly null than specific reference). On the other hand, this may also be a result of Polish politeness conventions: non-familiar interlocutors are supposed to be referred to in the third person. A Polish-Canadian speaker, who is aware of these conventions but who at the same time may not feel comfortable adhering to them, may make a compromise by using the second person with a null pronoun. Nonetheless, if we ignore the second person singular category, the order of factors within this factor group is the same for each generation. However, while for homeland speakers there is a dichotomy between first person tokens (66-77) and third person tokens (23-26), the same type of division does not exist for the two immigrant generations. In fact, for second generation speakers, the greatest difference is between third person singular and all other person+number combinations.

Second generation speakers also exhibit categorical behaviour with respect to tense. This may once again be due to the small number of future tense tokens since each generation exhibits very different behaviour for this tense. On the other hand, all generations agree that the fewest null subjects occur with verbs inflected for past feminine. Nonetheless, while homeland and first generation speakers treat present tense and past masculine tokens the same way, second generation speakers treat them differently so that past masculine tokens pattern more closely with past feminine tokens than with present tokens.

Inflection class is significant for both immigrant generations. However, the direction of factors within this factor group is very different depending on the generation. In fact, second

generation speakers follow the direction pattern of homeland speakers much more closely. The bar graph in Figure 5.1 below helps to visualize how first generation speakers stand apart from the other two generations for all inflection classes except (*ę ~ i*) and irregular (BE). Further note that second generation speakers are unique in favouring null subjects with verbs inflected using the (*ę ~ e*) ambiguous paradigm, as they are the only generation where the factor weight for this paradigm is above 0.5. Therefore, presence of phonetic ambiguity is clearly not linked to overt subject use for second generation speakers.

Figure 5.1: Distribution of null subjects based on generation and inflectional paradigm (N=934)



Finally, while the analysis in §5.2 shows that semantic content is significant for all data, when the data is divided by generation the factor is only significant for second generation speakers. Nonetheless, all generations agree that mental verbs disfavour null subjects.

There are also further differences among the generations with respect to factor groups not chosen as significant for any generation (Table 5.29).

Table 5.29

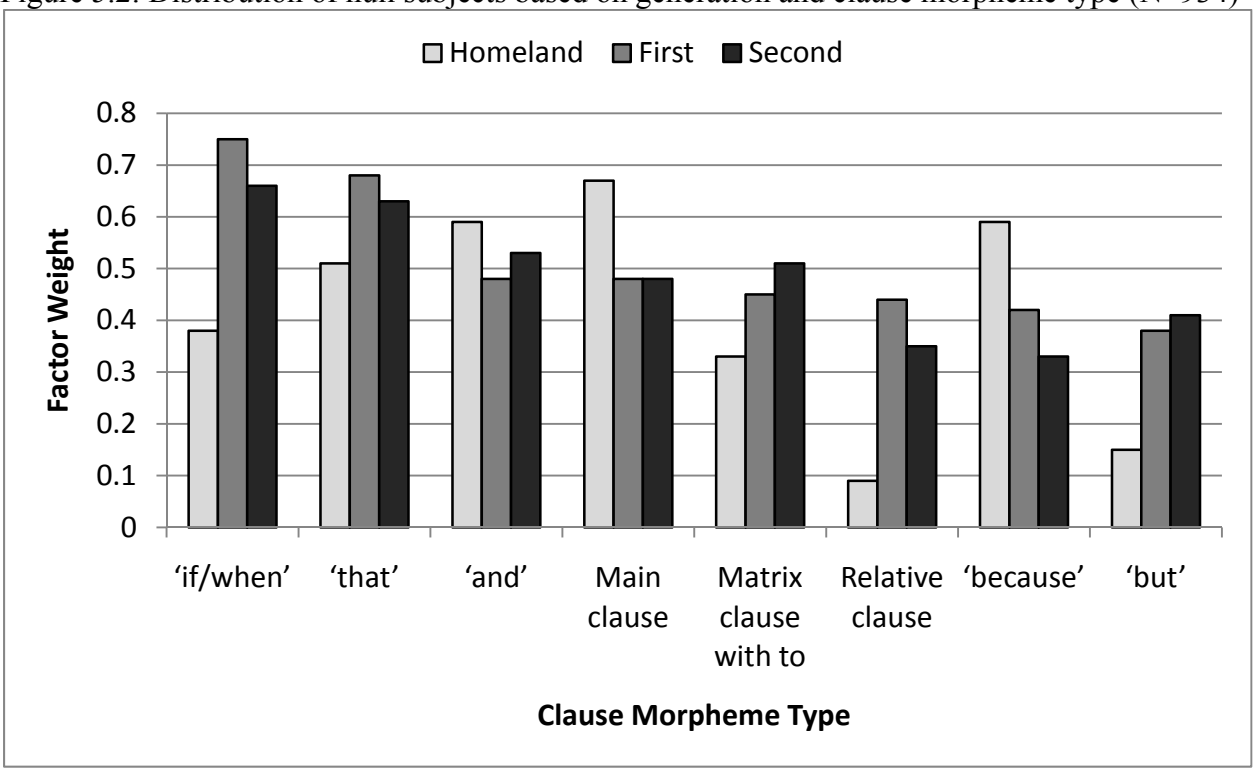
Comparison by generation of internal and external factors selected as *not significant* to the probability of Null subjects for any generation

	Homeland		First		Second	
Corrected mean:	0.908		0.833		0.779	
Total N	193		371		370	
	Factor weight	N	Factor weight	N	Factor weight	N
Clause type[6]						
'if/when'	[.38]	17	[.75]	28	[.66]	48
'that'	[.51]	14	[.68]	20	[.63]	26
'and'	[.59]	36	[.48]	47	[.53]	62
Main clause	[.67]	68	[.48]	177	[.48]	109
Matrix clause with *to*	[.33]	25	[.45]	35	[.51]	40
Relative clause	[.09]	5	[.44]	10	[.35]	13
'because'	[.59]	12	[.42]	26	[.33]	32
'but'	[.15]	16	[.38]	20	[.41]	40
Speaker sex						
Female	N/A		[.72]	54	[.49]	110
Male	N/A		[.46]	331	[.51]	265
Negation						
Positive	[.53]	181	[.50]	324	[.51]	321
Negative	[.18]	12	[.52]	47	[.43]	49

Specifically, recall from §5.2 that clause type was selected as significant in a multivariate analysis that included data from all speakers, yet it was not selected as significant for any of the generations individually. This may be due to fewer tokens in individual generations and the number of categories for the amount of data. However, the patterns show that there is no way we can further conflate categories and still be able to compare generations. As Figure 5.2 below shows, the two immigrant generations pattern very differently from homeland speakers.

[6] Some of the dramatic differences between generations for this factor could be attributed to the small number of tokens in any given category.

Figure 5.2: Distribution of null subjects based on generation and clause morpheme type (N=934)



Interestingly, while the two immigrant generations do not exhibit a big difference between main clauses and matrix clauses preceded by *to*, the homeland speakers do. Namely, main clauses are much more likely to occur with null subjects, while matrix clauses with *to* are much less likely to occur with null subjects.

5.4 Ethnic orientation analyses

An additional multivariate analysis run was done to include individual speakers as factors within one factor group. The results of this analysis returned the factor weights for each speaker (Table 5.30). This was meant to transform the raw null subject distribution into a revised null subject rate for each speaker (i.e. the null subject rate if all else in the data were equal among speakers). Therefore, the factor group was run with the other nine factors that were included in the multivariate analysis in §5.2.

Table 5.30

Distribution of Null subject variants by Speaker, from highest to lowest rate			
Corrected mean:			0.839
Total N			934
	Factor weight	%	N
Speaker			
2M23	.862	96.2	53
1M35	.739	90.7	54
0F16	.712	88.8	98
1F56	.673	86.8	53
2M29	.611	88.7	53
1M60	.579	84.6	52
1M55	.559	88.2	76
0F44	.520	72.6	95
2M21	.505	86.0	50
2M22	.481	84.9	53
2F18	.438	67.3	55
2F23	.350	68.5	54
1M88	.284	63.2	57
1M44	.277	62.0	79
2M47	.038	23.1	52

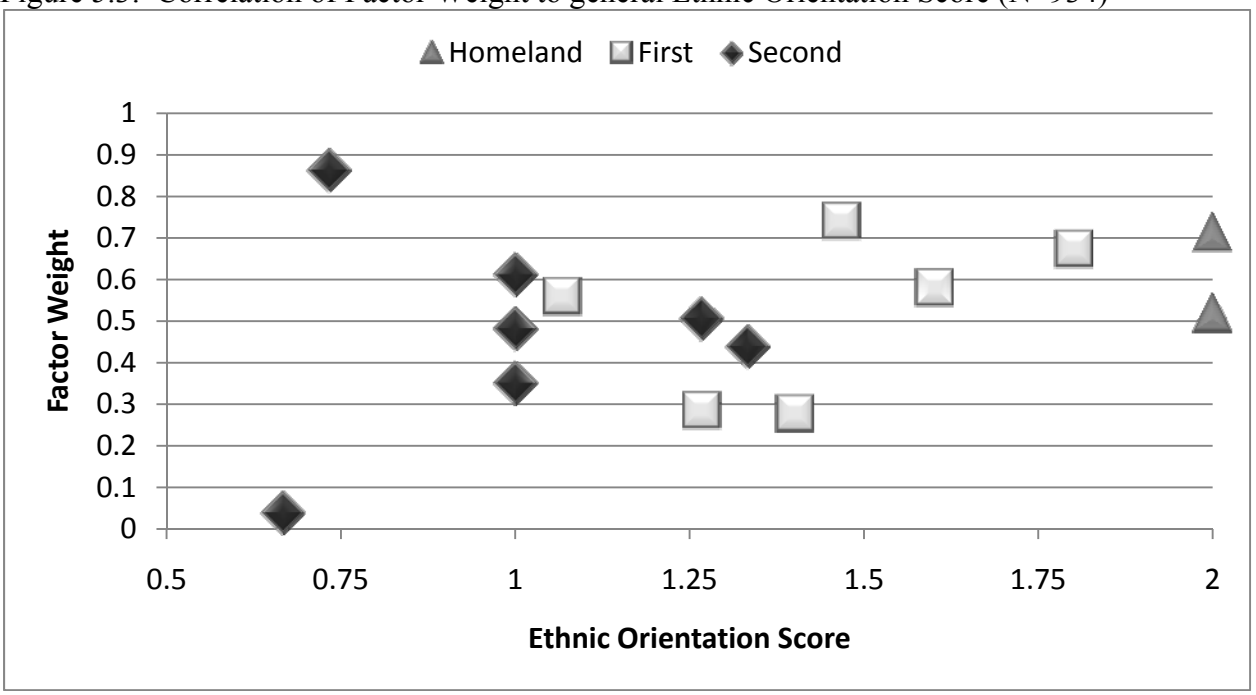
The data for the two homeland speakers shows that the null subject rate should be around .50 to .70. Anything outside this range is considered non-standard. The task is then to figure out the social factors behind why a speaker may have a rate that is significantly lower or higher.

5.4.1 Overall ethnic orientation

The factor weights for each speaker were plotted on a scattergram to determine the correlation between null subject rate and individual ethnic orientation. The overall individual ethnic orientation score was based on the reduced 15 measures of ethnic orientation (see Appendix E and F). This score places speakers on a continuum from being most

Canadian/English (0) to most Polish (2). In Figure 5.3, all homeland speakers are assigned an automatic ethnic orientation score of 2 for comparison purposes.

Figure 5.3: Correlation of Factor Weight to general Ethnic Orientation Score (N=934)



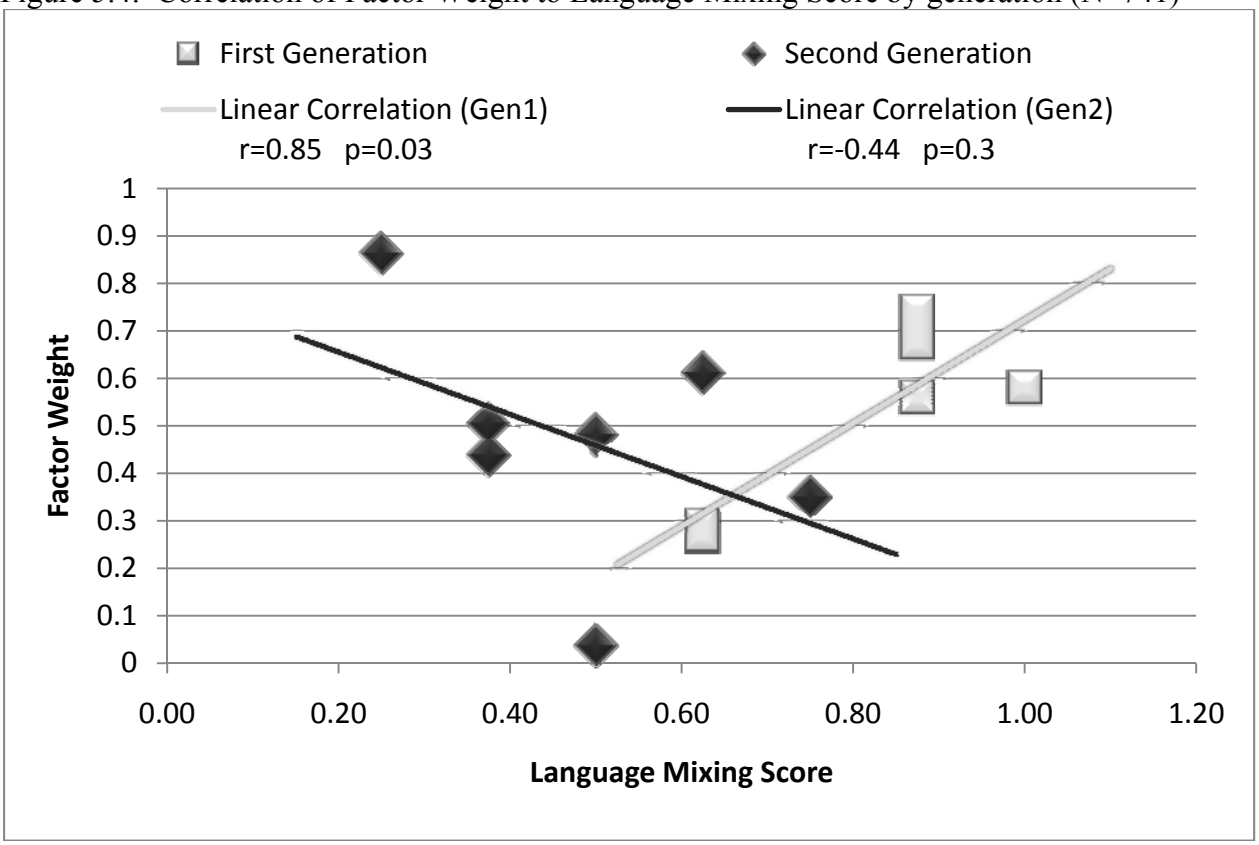
The scattergram in Figure 5.3 shows a lack of linear correlation between null subject rate and ethnic orientation. In fact, the two second generation speakers with the lowest ethnic orientation score are also the two speakers with the highest and lowest null subject rate. Furthermore, out of all the first generation speakers, the two with the lowest null rate fall somewhere in the middle with respect to ethnic orientation.

5.4.2 Language Mixing

However, there is a correlation between degree of language mixing (as defined in §4.1.2) and null subject use. This is particularly true for first generation speakers ($r=0.85$, $p=0.03$). Figure 5.4 shows that the two first generation speakers with the lowest null subject rate are also

the two speakers who tend to use both English and Polish in multiple social environments. On the other hand, those first generation speakers who have maintained a high null subject rate always or almost always use exclusively one language per speaking environment (see Appendix G).

Figure 5.4: Correlation of Factor Weight to Language Mixing Score by generation (N=741)



Note that the same type of correlation does not exist for second generation speakers. In fact, second generation speakers seem to exhibit a reverse correlation, with more language mixing resulting in a slightly higher null subject rate (r=-0.44, p=0.3).

5.4.3 Linguistic confidence

Although the null subject rate of first generation speakers correlates to language mixing, it does not correlate to linguistic confidence. However, linguistic confidence is very predictive of null subject rate when applied to second generation speakers ($r=0.98$, $p<0.01$).

Figure 5.5: Correlation of Factor Weight to Linguistic Confidence Score by generation (N=741)

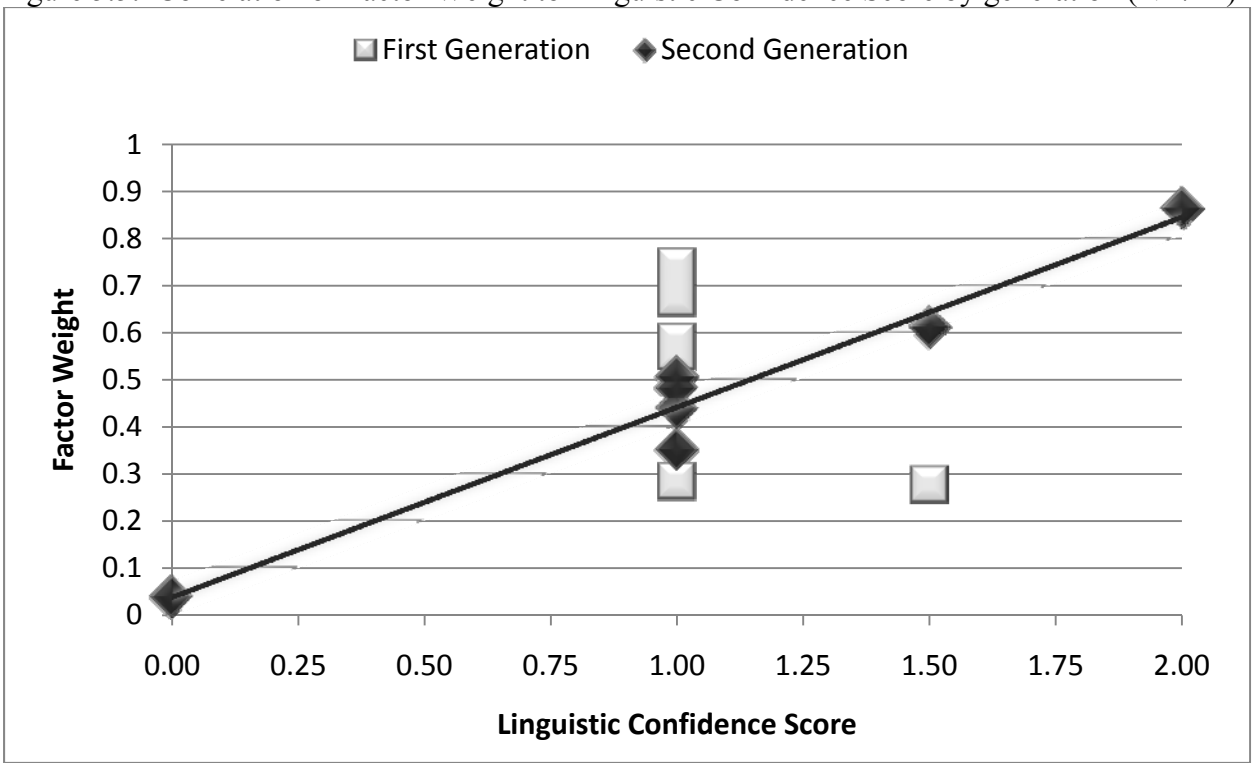


Figure 5.5 above shows that, in order to resist the influence of English on their null subject rate, speakers must have some desire to improve. That is, the more confident they are that their language is "good enough", the more likely they are to adopt English features, or simply to be unaware that they could be adopting English features. This also explains why one of the second generation speakers has a null subject rate that is even higher than the rates of homeland speakers. Speakers who are concerned about sounding too "Canadian" may exaggerate salient Polish features.

5.5 Summary of findings

Overall, the findings documented above show that Polish speakers in general exhibit a 77% null subject rate. The factors found significant for the entire Polish speaker sample are, from most significant to least significant: subject continuity, person and number subject agreement, immigrant generation, tense and gender inflection, semantic content of the lexeme, and the type of morpheme found in the clause head. On the other hand, inflection class, speaker sex, and presence of negation were not found to be significant.

Table 5.31

Summary of multivariate results with respect to hypotheses

Factor group	**Hypothesis**	**Supported by data?**
Person+number	1st person and plural subjects will be most likely null	**yes** most 1st person subjects are null
Tense+gender	past tense verbs with feminine gender marking will have most null subjects present tense & future perfect will have least null subjects	**no** past tense verbs with feminine gender marking have the least null subjects
Inflection class	*ę~e* paradigm will have significantly fewer null subjects	**no** factor group not significant
Semantic content	mental/estimative verbs will have fewest null subjects	**yes**
Subject continuity	same referents will have more null subjects	**yes**
Clause morpheme type	semantically contrastive morphemes will have fewer null subjects	**yes** 'but' has fewer null subjects than 'and'

Differences between generations were found to be significant. The null subject rate was found to decline with each subsequent generation. That is, the null subject rate is highest in homeland speaker, moderate in first generation speakers, and lowest in second generation speakers. Furthermore, not all generations exhibit the same type of behaviour with respect to

grammatical constraints affecting null subject use. Nonetheless, there are two factors that remain significant across all three generations: subject continuity and person+number agreement. However, the low significance of these two factors in second generation speakers leads me to conclude that while homeland speakers and first generation speakers have similar grammars, the grammar of second generation speakers is quite different.

Finally, social factor analyses found that a speaker's ethnic orientation alone cannot predict their null subject rate. Additional factors were found significant depending on the immigrant generation. Namely, language mixing was found significant for first generation immigrants, while linguistic confidence was found significant for second generation speakers.

6.0 DISCUSSION

6.1 Interpreting findings

6.1.1 Null subjects in Polish

Several aspects of null subject use in Polish follow the grammar of other null subject languages. For instance, the combination of person and number agreement on the verb was found to be one of the most significant factors. In particular, plural subject were more often null than singular subjects. This mirrors the patterns found in Italian (Nagy *et al.* 2010) and Spanish (Otheguy *et al.* 2007). Additionally, first person subjects were more often null than third person subjects, which in turn mirrors the patterns found in Cantonese (Nagy *et al*. 2010), Spanish (Otheguy *et al.* 2007), and English (Harvie 1998). Subject continuity was also found to be a significant factor, as has been the case for previous studies on all other languages including English. The fact that these variables are significant for so many languages, and that many languages follow similar patterns within these variables suggest that there are some universal properties to null subject use, regardless of language type.

Second, several popular assumptions about null subject use in Polish have been proven false. Polish speakers do not limit their use of overt pronouns to ambiguous contexts. This is true whether these contexts arise out of gender agreement ambiguity or phonetic ambiguity of verb forms. In fact, the results regarding gender agreement were completely contrary to the original hypothesis. Greater overt subject agreement on the verb, equated with the presence of gender agreement and feminine gender agreement in particular, was found to correlate with a decrease in null subject pronouns. Similarly, phonetically ambiguous verb forms, as illustrated by verbs belonging to the (*ę ~ e*) inflection class, are not different with respect to null subject pronouns than any other inflection class. That is to say, while verbs belonging to the strongly

unambiguous ($ę \sim {}^{i}e$) inflection class co-occurred with the most null subject pronouns, all other inflection classes behaved similarly with respect to each other. While strong lack of ambiguity causes an increase in null subject use, phonetic ambiguity does not decrease it.

Furthermore, there is some evidence that contrastive environments attract overt pronouns. For instance, out of all semantic content types, the lexemes with mental/estimative content were found to have the fewest null subjects. Similarly, verbs part of negative constructions and interrogative sentences were also found to have fewer null subjects than affirmative constructions and declarative sentences respectively. While these environments are not always necessarily contrastive, they are perhaps more likely to be contrastive than their counterparts. Another type of contrastive environment involves conjunctions meaning 'but'. When the subject continuity is different, these conjunctions occur with fewer null subjects than other types of conjunctions.

As suggested by the previous observation, while most clause morphemes behave similarly when the subject continuity is the same, they exhibit different patterns when the subject continuity is different. For instance, subordinate clauses headed by 'because' have significantly fewer null subjects; whereas parenthetical clauses such as those headed by 'when' or a relative morpheme behave the same regardless of subject continuity. The results for specific clause morphemes indicate that null subjects behave differently depending on the clause they occur in. However, if we were to conflate the clause morphemes into just three distinctions of main vs. coordinate vs. subordinate, no patterns would emerge. Therefore, it is important for future studies to pay attention to the particular, rather than the general, clause distinctions.

6.1.2 Changes across generations

The multivariate analyses performed on individual generations indicate a process of grammaticalization similar to that proposed by Heine & Kuteva (2005). At step one of Heine & Kuteva's process, a minor use pattern gains frequency; in other words, when applied to null versus overt subjects, the minor use patterns begins to replace the major use pattern. A comparison of speaker generations shows a gradual decrease in the frequency of null subject forms, with first generation speakers using fewer null subjects than homeland speakers, while second generation speakers use even fewer null subjects than first generation speakers. These results show that the minor use pattern is gaining frequency with each subsequent generation.

At step two, Heine & Kuteva predict that extension of the use pattern to new contexts will cause pragmatic factors such as subject continuity to lose significance. Once again, this is borne out in the data. While subject continuity has been identified as the most significant factor for both homeland speakers and first generation immigrants, it has been demoted to fifth most significant factor for second generation speakers.

At the third and final step of the grammaticalization model, it is expected that grammatical reanalysis will cause the use pattern to change meaning. In this case, the results show that a certain degree of grammatical reanalysis is already taking place in the first generation. For instance, while a combination of person and number has been found to be the second most significant factor group for both homeland and first generation speakers, first generation speakers have reanalyzed the significance of individual factors within this factor group. While homeland speakers show a dichotomy between all third person and all first person tokens, first generation speakers do not. Furthermore, while homeland speakers find tense to be a significant

factor, it is no longer significant for first generation speakers, who instead find inflection class to be significant.

Second generation speakers exhibit even more divergence from homeland speakers than first generation speakers. While they continue to find person+number significant, this time the speakers reanalyze it to include a dichotomy between third person singular tokens versus everything else. Similarly, while second generation speakers reinstate tense+gender as significant, they treat present tense forms and past masculine forms differently, unlike homeland speakers for whom the two factors are the same.

In many ways, second generation speakers pattern as a combination of first generation and homeland speakers. For instance, while second generation speakers resemble first generation immigrants in that they find inflection class to be significant, the ordering of factors within this factor group is more similar to that found in homeland speakers. Finally, second generation speakers are unique in finding semantic content a significant—in fact the most significant—factor.

These results all show that the grammaticalization process proposed by Heine & Kuteva (2005) is borne out in the Polish data. However, the steps in this process may not occur as discretely as that model proposes. Rather, they seem to overlap and move forward simultaneously.

6.1.3 Influence of ethnic orientation

The results of correlating each speaker's null subject rate with their overall ethnic orientation score showed that ethnic orientation alone neither attracts nor resists influence from

English. Language mixing measures correlate to null subject use in first generation speakers, while linguistic confidence measures correlate to null subject use in second generation speakers.

The language mixing data showed that simply speaking English does not lead to increased presence of English features in Polish. Rather, such influence emerges out of language mixing, as evidenced by the ability or option of using both languages in the same environment. These findings may also explain an unexpected result in Otheguy *et al.*'s (2007) study where increased use of Spanish in the workplace was found to positively correlate with overt subject use. Because many workplaces in America use English to some degree (to communicate with supervisors, coworkers of a different ethnicity, or clients), an increased use of Spanish may also indicate increased language mixing. If the Spanish speakers surveyed in Otheguy *et al.*'s study do in fact work in non-exclusively Spanish work environments, it is not surprising that English is more likely to influence those who speak Spanish at work. On the other hand, if the two languages are kept separate, they are also more likely to remain separate with respect to grammatical features.

The fact that language mixing fails to correlate in the same way for second generation speakers is not surprising either; the result of being born and growing up in an English speaking country means that most environments are linguistically mixed for second generation speakers. For most second generation speakers, there are very few environments which are exclusively Polish. Therefore, a predominance of language mixing environments may lead to an awareness of the difference between the two languages, and perhaps the desire to differentiate the two languages as much as possible. This, in conjunction with the findings for linguistic confidence measures, would explain the opposite correlation of null subjects to language mixing measures for second generation speakers.

The results for linguistic confidence in second generation speakers shows that those speakers who do not feel confident in how well or how often they speak Polish are more likely to have a high null subject rate than those speakers who are less concerned about their language use. Therefore, speakers who are worried that English may be influencing how they speak Polish will try to resist these influences. This explains the inflated use of null subjects by the one speaker (2M23) who has both the lowest language mixing score, and the highest linguistic confidence score. The results for this one speaker, whose null subject rate is even higher than that reported for homeland speakers, is further supported by Bayley & Pease-Alvarez's (1997) results for Spanish speakers of Mexican descent living in California. Spanish speakers whose mothers were born in the U.S. or left Mexico as young children have higher null subject rates than those speakers whose mothers left Mexico as young adults or who were themselves born in Mexico. That is, speakers with a greater depth of ties to the U.S. may be more concerned about whether they are speaking "proper" Spanish, and may therefore exaggerate salient features of their heritage language.

Given these findings, it is not surprising that the null subject rates of first generation speakers do not correlate to linguistic confidence. Since these speakers were born and raised in Poland and since most of them consider Polish to be their primary language, they are generally confident in their use of Polish. In my experience, people who moved to Canada as adults do not worry about whether they are "forgetting" how to speak Polish; they are more concerned about learning English. In fact, they do not believe that it is even possible for them to "forget" the language that they have used exclusively while growing up.

6.2 Suggestions for future research

The above discussion about ethnic orientation suggests more research needs to be done in the field to determine the relevance and implication of each measure. For instance, I believe the linguistic confidence measure I devised may be purely related to ethnic orientation. That is, the desire to disassociate oneself from a mixed culture may be an important factor. Those speakers who have inflated null subject rates will be positively oriented towards Polish, and may be positively oriented towards English, but will be negatively oriented towards "Polonia" (i.e. Polish-Canadian) culture. Therefore, it is important that future studies gather information about a speaker's ethnic orientation and attitudes regarding the three types of cultural distinctions above.

Furthermore, the current linguistic corpus will need to be expanded; more speakers need to be interviewed in order to strike a balance in the speaker distribution. This includes interviewing more homeland speakers, as well as including a wider variety of speakers in terms of sex and age. In particular, it will be important to interview homeland males and additional first generation females to determine whether sex is a factor in null subject use. Also, additional older second generation speakers will need to be interviewed to facilitate better comparisons to first generation speakers. Moreover, a larger speaker sample will help confirm the ethnic orientation findings.

Extracting additional tokens from the data will also be necessary to examine patterns within rare factors. For instance, due to the small number of tokens, it was not possible to examine patterns with respect to second person singular tokens. This included being able to distinguish these tokens based on specific versus generic reference. Similarly, the scope of this paper did not allow for a thorough examination of subject continuity. Future statistical analyses

that divide tokens based on the syntactic position of their antecedents will be necessary. The significance of such an approach has been suggested by the results for tokens whose subjects are coreferential with the object of a preceding ergative construction.

Other findings in the current study hint at the necessity for coding additional factors. For instance, the low null subject rate found for the verb BE in comparison to all other verbs suggests a significance of lexical frequency. Accordingly, future research should examine the frequency of lexemes with respect to null subject rate. Similarly, while past tense feminine tokens were found to occur with the fewest null subjects, it is not certain whether this is only due to the overt feminine marking on the verb. It may be necessary to examine whether the inherent grammatical gender of the subject plays a role, regardless of whether or not the verb is inflected for gender agreement.

Finally, additional variable contexts that did not fall into the current envelope of variation will need to be studied. This includes imperative constructions (which are always null in English, but are supposedly variable in Polish), and verbs with the generic subject *to* 'it' (which may be null). The inclusion of inherently subject-less constructions in the envelope of variation may also prove interesting. For instance, some speakers who have a particularly low null subject rate may also use more subject-less constructions, resulting in a high null subject rate overall. I speculate that speakers who have had more contact with English will be less likely to use the (uniquely Polish) subject-less constructions.

7.0 CONCLUSIONS

Contrary to popular belief, it is not true that overt subjects in Polish are emphatic, contrastive, or otherwise marked. While contrastive environments such as changes in subject continuity, questions, negation, and contrastive conjunctions tend to have a lower null subject rate, none of these patterns are categorical. Similarly, ambiguity is not a significant factor, neither with respect to phonetic morphology, nor richness of agreement. Overt subjects still appear in non-contrastive and non-ambiguous environments.

Nonetheless, the data supports the hypothesis that certain universal patterns with respect to subject continuity exist. Namely, it supports the finding that subject continuity and person+number agreement are significant factors in the distribution of null subjects in most languages. This not only includes typical null subject languages such as Spanish (Otheguy *et al*. 2007), Italian (Nagy *et al*. 2010), and Russian (Hollett 2010), but also includes canonically overt subject languages such as English (Harvie 1998).

Overall this indicates that there are general pragmatic reasons that drive null subject use. Thus, verbs whose subject is correferential with the one that came before significantly favour null subjects to avoid redundancy, while verbs whose subject is not correferential with the one that came before significantly disfavour null subjects to signal a switch in topic. Similarly, it can be said that, because humans tend to talk about themselves and their own experiences most often, the first person is the most frequent subject. Thus, first person subjects significantly favour the null form because they are the assumed subject in most contexts. On the other hand, factors related to the particular form of the verb are not significant. As a result, it is not surprising that Polish past tense forms disfavour null even though they have greater subject agreement. Nor is it surprising that the ambiguous inflection class favours null for second generation speakers.

Neither tense nor inflection class are dependent on the context. Therefore, context, as opposed to form, contributes more to our ability to reconstruct verb subject referents. This may explain the presence of null subjects in non-inflecting languages such as Cantonese.

In this respect, because all three immigrant generations of Polish speakers significantly make use the same two universal factors (switch reference and grammatical person) in deciding between null and overt subjects, their grammars are essentially the same. Nonetheless, the process of grammatical replication proposed by Heine & Kuteva (2005) has been borne out in the data. The minor use pattern involving overt subjects is gaining in frequency with each subsequent immigrant generation, at the expense of null subject use. Second, a pragmatic factor (subject continuity) has largely lost significance for second generation speakers. Third, a type of grammatical reanalysis is taking place in the second generation: new constraints are gaining significance. Overall, steps are being taken in changing the canonical null subject language status of Heritage Polish. However, these generational changes are curious considering that English, a canonically overt subject language, employs the same universal constraints on null subjects as other canonically null subject languages, including homeland Polish. Therefore, it is not that Heritage Polish is directly moving towards an English grammar; rather, it is moving away from Homeland Polish.

Finally, the ethnic orientation data shows that factors other than contact with English alone play a role in the degree and direction of change. The results of measuring language mixing and linguistic confidence indicate that traditional ways of coding ethnic orientation must be revised. We cannot simply place speakers on a continuum from Language A to Language B, nor can we take answers at face value without also asking speakers to qualify what these answers mean to them.

REFERENCES:

Bartmiński, Jerzy (ed.) 2001. *Współczesny język polski [Contemporary Polish]*. Lublin: Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej.

Bayley, Robert and Lucinda Pease-Alvarez. 1997. Null pronoun variation in Mexican-descent children's narrative discourse. *Language Variation and Change* 9, 349-371.

Bolonyai, Agnes. 2000. "Elective affinities": Language contact in the Abstract Lexicon and its structural consequences. *International Journal of Bilingualism* 4:1, 81-106.

Bondaruk, Anna. 2001. Pro-drop in Polish and Irish. *Poznań Studies in Contemporary Linguistics* 37, 35-52.

Dubisz, Stanisław (ed.) 1997. *Język polski poza granicami kraju*. Opole : Uniwersytet Opolski-Instytut Filologii Polskiej.

Frajzyngier, Zygmunt & Erin Shay. 2003. *Explaining language structure through systems interaction*. Philadelphia PA: John Benjamins.

Gruszczyński, Włodzimierz & Jerzy Bralczyk (Eds). 2002. *Słownik gramatyki języka polskiego [Dictionary of Polish grammar]*. Warsaw: Wydawnictwa Szkolne i Pedagogiczne.

Harvie, Dawn.1998. *Null* subject in English: Wonder if it exists? *Cahiers Linguistique d'Ottawa* 26, 15–25.

Heine, Bernd & Tania Kuteva. 2005. *Language Contact and Grammatical Change*. Cambridge University Press.

Hoffman, Michol & James Walker. 2010. Ethnolects and the city: Ethnic orientation and linguistic variation in Toronto English. *Language Variation and Change* 22, 37-67.

Hollett, Meghan. 2010. Heritage Russian in Toronto: Generational change and subject omission. Paper presented at *CVC IV*, Memorial University of Newfoundland, June 19-20 2010.

Janicki, Karol. 1995. Sociolinguistics in Poland: History and prospects (A socio-political analysis). In *When East met West : sociolinguistics in the former socialist bloc*, Jeffrey Harlig & Csaba Pléh (eds.) Berlin: Mouton de Gruyter.

Johanson, Lars. 2008. Remodeling grammar: Copying, conventionalization, grammaticalization. In *Language Contact and Contact Languages*, Peter Siemund & Noemi Kintana. Amsterdam: John Benjamins.

Keefe, Susan & Amado Padilla. 1987. *Chicano Ethnicity*. Albuquerque, NM: University of New Mexico Press.

Lindseth, Martina. 1998. *Null-subject properties of Slavic languages : with special reference to Russian, Czech, and Sorbian*. Munich: Otto Sagner.

Lipski, John M. 1996. Patterns of Pronominal Evolution in Cuban-American Bilinguals. In *Spanish in Contact: Issues in Bilingualism* Ana Roca & John B. Jensen (eds.). Somerville: Cascadilla, 159-186.

Lubaś, Władysław. 2003. *Polskie gadanie: podstawowe cechy i funkcje potocznej odmiany polszczyzny [Polish talk: basic characteristics and functions of the colloquial variety of Polish]*. Opole : Wydawn. Uniwersytetu Opolskiego

Lustanski, Joanna. 2009a. Polish Canadians and Polish immigrants in Canada: self-identity and language attitude. *International Journal of the Sociology of Language* 199, 39-61.

Lustanski, Joanna. 2009b. *Jezyk polonijny w Kanadzie [Spoken Canadian Polish]*. Toronto: The Canadian Polish Research Institute.

McShane, Marjorie. 2009. Subject ellipsis in Russian and Polish. *Studia Linguistica* 63:1, 98-132.

McShane, Marjorie. 2005. *A theory of ellipsis*. Oxford: Oxford University Press.

Mesthrie, Raj & Rakesh Bhatt. 2008. *World Englishes: The Study of new linguistic varieties*. Cambridge University Press.

Miodunka, Władysław T. 2003. *Bilingwizm Polsko-Portugalski w Brazylii: w stronę lingwistyki humanistycznej [Polish-Portuguese Bilingualism in Brasil]*. Kraków: Universitas.

Montrul, Silvina A. 2004. Subject Expression in Spanish Heritage Speakers: A Case of Morphosyntactic Convergence. *Bilingualism: Language and Cognition* 7, 125-142.

Morales, Amparo. 1997. La hipotesis funcional y la aparicion del sujeto no nominal: El espanol de Puerto Rico. *Hispania* 80, 153–65.

Myers-Scotton, Carol. 2002. *Contact Linguistics: Bilingual Encounters and Grammatical Outcomes*. Oxford: Oxford University Press.

Nagy, Naomi, Nina Aghdasi, Derek Denis, Alexandra Motut & Dylan Uscher. 2010. Pro-drop in Heritage Languages: A cross-linguistic study of contact-induced change**.** Paper presented at *NWAV 39*, San Antonio, TX.

Niedzielski, Henry. 1997. First-language maintenance among twentieth century Polish immigrants to France, the United States, and New Zealand. In *Language history and linguistic modelling : a festschrift for Jacek Fisiak on his 60th birthday (vol.2)*, Raymond Hickey and Stanisław Puppel (eds.) Berlin: Mouton de Gruyter, 1765-1780.

Otheguy *et al.*, Ricardo, Ana Celia Zentella, and David Livert. 2007. Language and Dialect Contact in Spanish in New York: Toward the Formation of a Speech Community. *Language* 83:4.

Polinsky, Maria. 1995. Cross-linguistic parallels in language loss. *Southwest Journal of Linguistics* 14:1-2, 87-123.

Savić, Jelena M. 1995. Structural convergence and language change: Evidence from Serbian/English code-switching. *Language in Society* 24:4, 475-92.

Schmitt, Elena. 2000. Overt and covert codeswitching in immigrant children from Russia. *Int'l Journal of Bilingualism* 4,9-28.

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

Statistics Canada. 2006. Population by mother tongue, by census metropolitan area. <http://www40.statcan.gc.ca/l01/cst01/demo12c-eng.htm>

Toribio, Almeida Jacqueline. 2004. Convergence as an Optimization Strategy in Bilingual speech: Evidence from Code-Switching. *Bilingualism: Language and Cognition* 7, 165-173.

Torres Cacoullos, Rena, and Catherine E. Travis. 2010. Variable *yo* expression in New Mexico: English influence? *Spanish of the U.S. Southwest: A language in transition*, Susana Rivera-Mills & Daniel Villa (eds.). Frankfurt: Iberoamericana/Vervuert, 185-206.

Travis, Catherine E. 2007. Genre Effects on Subject Expression in Spanish: Priming in Narrative and Conversation. *Language Variation and Change* 19, 101-135.

APPENDIX A: Maps

Figure A-1 shows a map of Poland with current voivodeship (province) borders marked. Shaded voivodeships indicate areas where speakers in this study originated. Either they grew up there, or their parents did. Speaker codes of second generation speakers whose mother did not originate from the same area as their father are written in brackets. Note that one speaker's parent (2M47) has origins in Belarus (which used to be a part of Poland).

Figure A-1

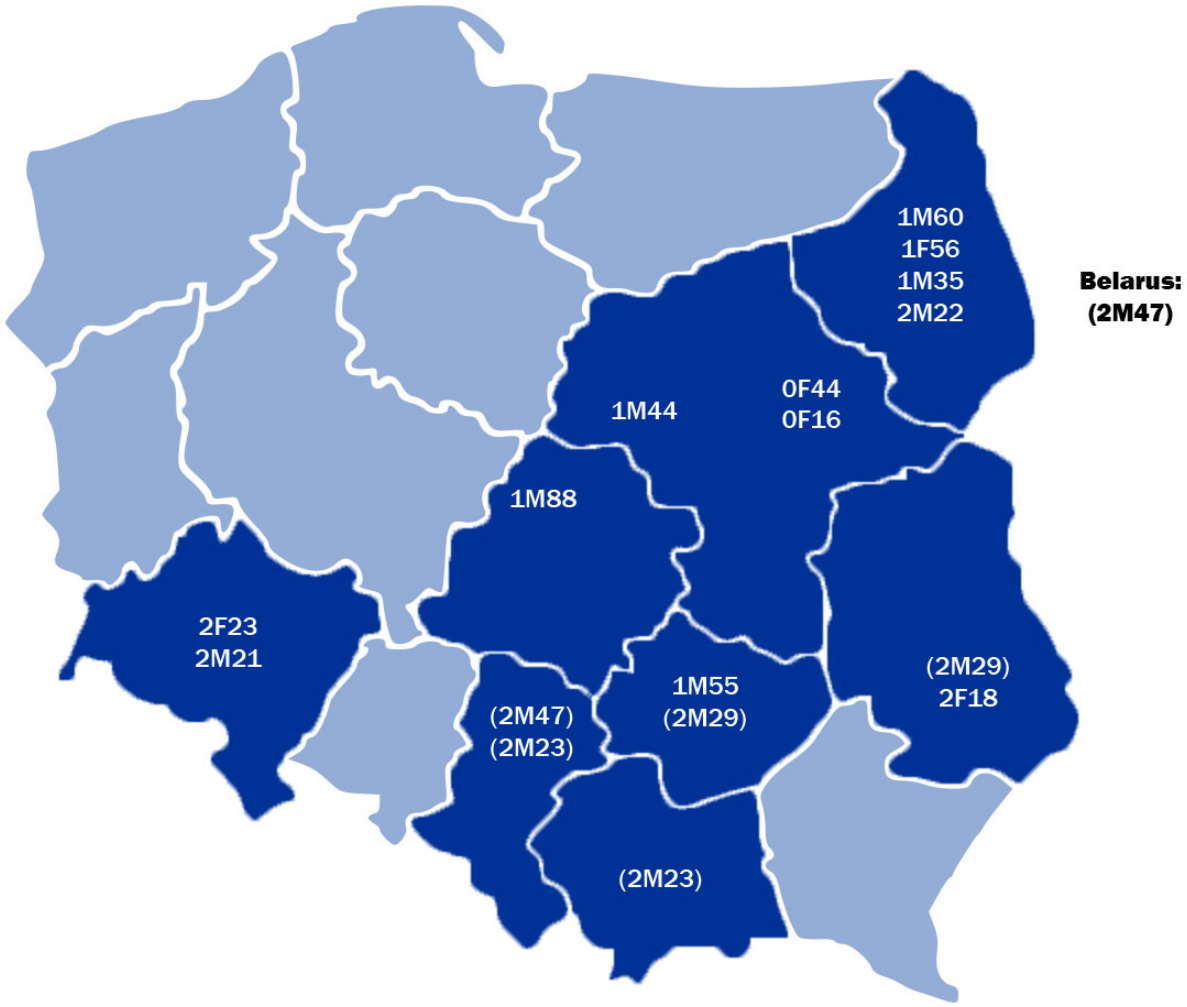
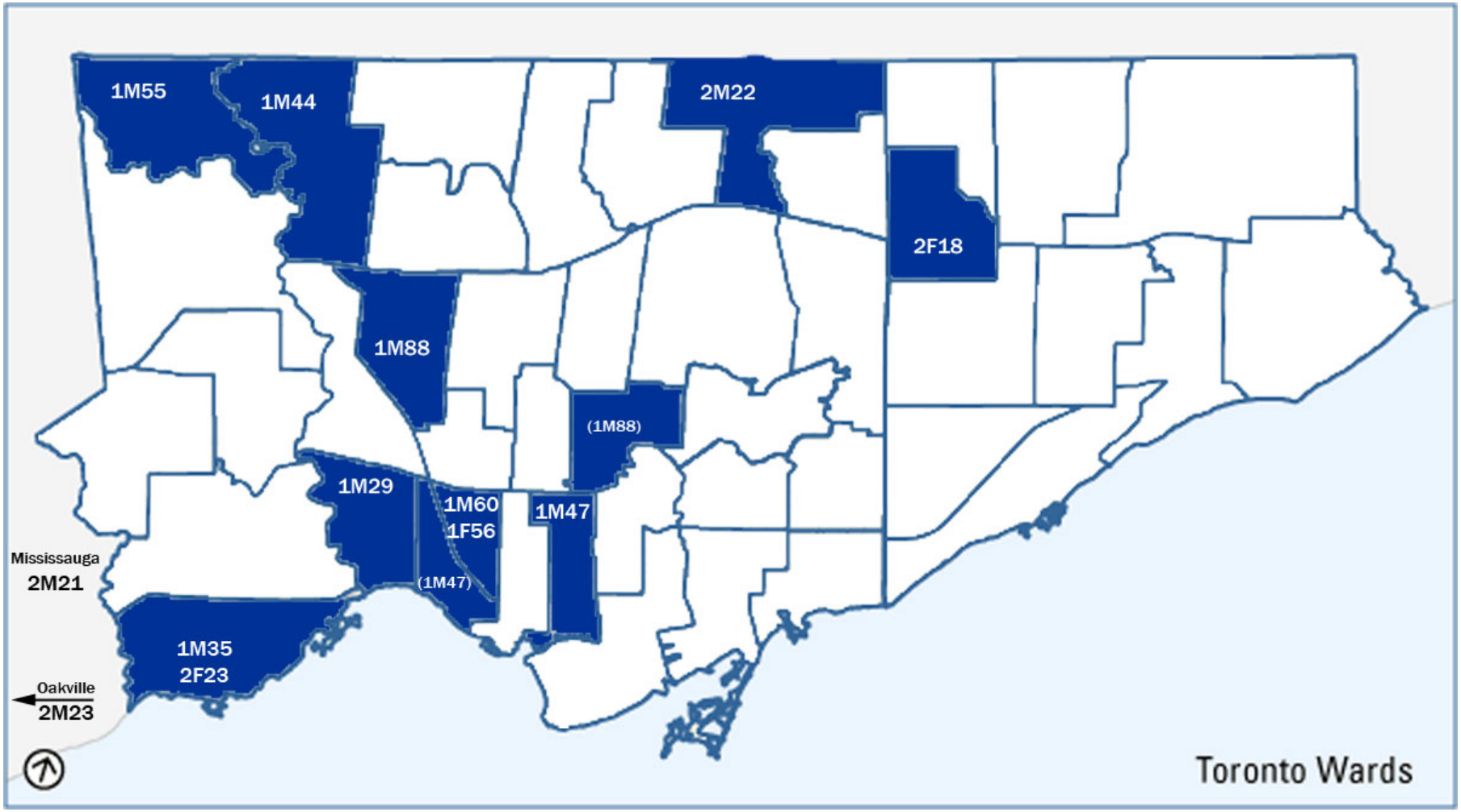


Figure A-2 shows a map of Toronto with ward (neighbourhood) borders marked. Shaded wards indicate areas where speakers in this study live. Speaker coded listed in brackets indicate that in the past the speaker used to live in that neighbourhood for a significant length of time.

Figure A-2


1M55
1M44
2M22
2F18
1M88
(1M88)
1M29
1M60
1F56
1M47
(1M47)
Mississauga
2M21
1M35
2F23
Oakville
2M23
Toronto Wards

APPENDIX B: Factor by factor distribution by immigrant generation

Table B-1:

Distribution of Null variants by grammatical person and number agreement

	Homeland		First		Second	
Person	%	N	%	N	%	N
1 sg.	87	67	82	191	74	273
2 sg.	85	13	83	23	100	13
3 sg.	65	55	67	66	44	18
1 pl.	92	38	85	47	85	123
3 pl.	80	20	70	44	68	28
Total N	193		371		370	

Table B-2:

Distribution of Null variants by tense

	Homeland		First		Second	
Tense	%	N	%	N	%	N
Future	94	17	81	16	100	9
Present	86	74	77	185	77	191
Past masculine	84	45	82	148	71	119
Past feminine	67	57	64	22	61	51
Total N	193		371		370	

Table B-3:

Distribution of Null variants by presence of negation

	Homeland		First		Second	
Negation	%	N	%	N	%	N
Positive	81	181	78	324	74	321
Negative	75	12	79	47	69	49
Total N	193		371		370	

Table B-4:

Distribution of Null variants by semantic content of the verb

	Homeland		First		Second	
Semantic content	%	N	%	N	%	N
Action	81	104	80	196	81	167
Stative	82	61	78	115	72	133
Mental	79	28	75	60	57	70
Total N	193		371		370	

Table B-5:

Distribution of Null variants by inflection class

	Homeland		First		Second	
Inflection	%	N	%	N	%	N
irregular	81	27	68	31	65	52
ę ~ e	76	25	71	42	76	58
ę ~ i	82	56	80	104	69	80
ę ~ ie	89	36	79	43	93	41
am ~ a	74	42	84	138	70	118
em ~ e	86	7	54	13	86	21
Total N	193		371		370	

Table B-6

Distribution of Null variants by subject continuity

	Homeland		First		Second	
Subject continuity	%	N	%	N	%	N
Same	91	101	87	215	79	230
Different	70	92	67	156	65	140
Total N	193		371		370	

Table B-7:

Distribution of Null variants by clause morpheme

	Homeland		First		Second	
Clause morpheme	%	N	%	N	%	N
‘if/when’	82	17	89	28	88	48
‘that’	86	14	85	20	81	26
to	72	25	80	35	80	40
Main	90	68	77	177	72	109
‘and’	81	36	79	47	76	62
Relative clause	60	5	80	10	62	13
‘because’	75	12	73	26	56	32
‘but’	62	16	71	20	65	40
Total N	193		371		370	

APPENDIX C: Original ethnic orientation questionnaire

The following outlines the original 37-question ethnic orientation questionnaire (from Hoffman & Walker 2010, based on Keefe & Padilla 1987) that formed the basis of the sociolinguistic interviews used in this study. Following each question, possible answers are listed alongside the number of points that would be assigned to a speaker who had given such an answer.

Ethnic identification:

1. Do you think of yourself as Polish, Canadian or Polish-Canadian?
   0 Canadian
   1 Polish-Canadian
   2 Polish
2. Are most of your friends Polish?
   0 No
   1 mixed
   2 Yes
3. Are people in your neighbourhood Polish?
   0 No
   1 mixed or used to be
   2 Yes
4. Are the people you work with Polish?
   0 No
   1 mixed
   2 Yes
5. When you were growing up, were the kids in your school Polish? Were your friends? The kids in your neighbourhood?
   0 No
   1 mixed
   2 Yes

Language:

1. a. How well do you speak Polish?
      0 Not well at all
      1 Admits to making major mistakes
      2 Well
   b. How often do you speak Polish?
      0 Never
      1 Sometimes
      2 Everyday

2. Where did you learn Polish? At home? In school?
   0 School
   1 both
   2 Home
3. Do you prefer to speak Polish or English?
   0 English
   1 both
   2 Polish
4. a. Do you prefer to read and write in Polish or English?
      0 English
      1 both
      2 Polish
   b. Do you read Polish magazines and newspapers? — Which ones?
      0 no
      2 yes
5. Do you prefer to listen to the radio or watch TV in Polish or English?
   0 English
   1 both
   2 Polish

Language choice:

1. What language does your family speak when you get together?
   0 English
   1 both
   2 Polish
2. What language do you speak with your friends?
   0 English
   1 both
   2 Polish
3. What language do you speak when you're talking about something personal? When you're angry?
   0 English
   1 both
   2 Polish
4. Did/do you speak to your parents in Polish? Your grandparents?
   0 English
   1 both
   2 Polish
5. Do you speak to your children/grandchildren in Polish?
   0 English
   1 both
   2 Polish

Cultural heritage:

1. a. Where were you born?
   0 Canada
   1 somewhere else: Insert a "comment" to specify place.
   2 Poland
   b. If in Poland: How old were you when you came here?
   1 less than 18
   2 18 or older
   c. If in Canada: Have you ever been to Poland? When? For how long?
   0 has never been to Poland
   1 visited
   2 lived there for more than 3 months at a stretch
2. Where did you go to school?
   0 Canada
   1 somewhere else
   2 Poland

Parents:

1. Do your parents think of themselves as Polish, Canadian or Polish-Canadian?
   0 Canadian
   1 both or mixed (1 parent each way, or Polish-Canadian)
   2 Polish
2. a. Do/did your parents speak Polish? English?
   0 English
   1 both (one parent each, or at least one who spoke both)
   2 Polish
   b. Do/did your grandparents speak Polish? English?
   0 English
   1 both
   2 Polish
3. a. How old were your parents when they came to Canada?
   0 born in Canada
   1 <18
   2 18+
   b. How old were your parents when they came to Canada?
   0 born in Canada
   1 <18
   2 18+

Partner:

1. Is your husband/wife/boyfriend/girlfriend Polish?
   0 No
   1 Yes, at least part-Polish, but born in Canada
   2 Yes. Born in Poland

2. Does she/he think of her/himself as Polish, Canadian or Polish-Canadian?
   0 Canadian or some other "hyphenated-Canadian"
   1 Polish-Canadian
   2 Polish
3. Does she/he speak Polish? Do you speak Polish to her/him?
   0 No
   1 Yes she/he can, but they don't (or VERY rarely).
   2 Yes, they speak Polish, at least sometimes.

Polish culture:

1. Should Polish-Canadian kids learn Polish? Polish culture?
   0 No
   1 don't know / maybe
   2 Yes
2. Would you rather live in a Polish neighbourhood?
   0 No
   1 don't know / maybe
   2 Yes
3. Should Poles only marry other Poles?
   0 No
   1 No, but it should be someone of a "similar type"
   2 Yes

Discrimination:

1. Have you ever had a problem getting a job because you're Polish?
   0 No
   2 Yes
2. What about renting an apartment or buying a house?
   0 No
   2 Yes
3. Were you treated differently by your teachers in school?
   0 No
   2 Yes
4. Have you ever been treated badly because you're Polish?
   0 No
   1 nothing serious
   2 Yes
5. Is there a lot of discrimination against Polish people?
   0 No
   1 aware of historical discrimination or stereotypes
   2 Yes

APPENDIX D: Answers to original ethnic orientation questionnaire

The following table summarizes the answers each speaker gave in response to the original ethnic orientation questionnaire (see Appendix C). Speaker codes are given in the leftmost column; note that data from speakers P1F56 and P1M35 is absent since they were not asked to respond to the original questionnaire.

At the top of each column, each question is identified by its corresponding code. For example, "A1" corresponds to the first question in section A of the questionnaire: "*Do you think of yourself as Polish, Canadian or Polish-Canadian?*"

Points were assigned based solely on the answer given by a speaker during the interview. An "x" indicates that an answer to a particular question is not available for that speaker.

Finally, at the end of the table, the total number of points for each particular speaker is tabulated. This total is then divided by the total number of questions that the speaker has answered to give the average orientation score.

Speaker	Ethnic Identification					Language						
	A1	A2	A3	A4	A5	B1a	B1b	B2	B3	B4	B4b	B5
P1M88A	x	1	0	x	2	2	2	2	x	2	2	x
P1M60A	2	2	0	0	2	2	2	2	2	2	1	1
P1M55A	1	0	0	0	x	2	1	x	2	2	x	0
P1M44A	2	1	1	2	x	2	2	x	1	2	2	1
P2M47A	1	1	1	0	2	1	1	2	0	0	x	0
P2M29A	1	0	0	0	1	2	2	2	1	0	2	0
P2M23A	1	0	0	2	0	2	1	2	1	1	1	0
P2M22A	1	1	0	0	0	2	2	2	0	0	1	1
P2M21A	1	1	1	0	1	2	2	2	2	1	2	2
P2F23A	1	0	2	0	1	2	2	2	1	0	0	0
P2F18A	1	1	1	2	1	2	2	2	1	0	2	1

Speaker	Language Choice					Cultural Heritage			
	C1	C2	C3	C4	C5	D1a	D1b	D1c	D2
P1M88A	1	1	x	x	0	2	2	0	x
P1M60A	2	2	2	x	2	2	2	1	2
P1M55A	x	x	x	x	x	2	2	1	x
P1M44A	x	x	x	x	x	2	2	x	x
P2M47A	1	1	x	1	x	0	0	1	0
P2M29A	2	0	2	2	x	0	0	1	1
P2M23A	2	1	1	2	x	1	0	1	0
P2M22A	2	1	0	2	x	2	1	x	0
P2M21A	2	1	1	2	x	0	0	1	1
P2F23A	2	0	1	2	x	2	1	2	1
P2F18A	2	1	1	2	x	0	0	1	1

Speaker	Parents					Partner			Polish Culture		
	E1	E2a	E2b	E3a	E3b	F1	F2	F3	G1	G2	G3
P1M88A	x	x	x	x	x	0	0	0	x	x	x
P1M60A	x	x	x	x	x	2	2	2	2	0	0
P1M55A	x	x	x	x	x	0	0	0	2	x	x
P1M44A	x	2	x	x	x	0	0	0	2	x	x
P2M47A	x	1	x	x	x	x	x	x	2	x	x
P2M29A	2	2	x	2	x	x	x	x	2	2	0
P2M23A	2	1	1	2	x	0	x	1	2	1	0
P2M22A	2	2	x	2	x	1	x	x	2	2	1
P2M21A	2	1	x	2	x	1	x	x	2	1	1
P2F23A	1	1	2	2	x	2	2	2	2	1	0
P2F18A	2	1	2	2	x	x	x	x	2	1	1

Speaker	Discrimination					Total	Average
	H1	H2	H3	H4	H5		
P1M88A	x	x	x	x	x	19	1.12
P1M60A	0	0	0	0	0	41	1.32
P1M55A	x	x	x	1	1	17	0.94
P1M44A	x	x	x	x	2	26	1.44
P2M47A	x	x	x	x	2	18	0.86
P2M29A	1	0	0	2	1	33	1.06
P2M23A	0	0	0	0	1	30	0.88
P2M22A	0	0	0	0	1	31	1.00
P2M21A	0	x	0	0	1	36	1.16
P2F23A	0	0	0	0	0	37	1.06
P2F18A	0	0	1	1	1	38	1.19

APPENDIX E: Revised ethnic orientation measures

The following outlines the revised 15 ethnic orientation measures. Following each measure, possible statements are listed alongside the number of points that would be assigned to a speaker who had made such a statement.

Ethnic Identity

1. Speaker
   - *2 = Polish*
   - *1 = Polish-Canadian*
   - *0 = Canadian*
2. Speaker's parents
   - *2 = Polish*
   - *1 = Polish-Canadian*
   - *0 = Canadian*
3. Speaker's partner
   - *2 = Polish*
   - *1 = had been with a Polish person in the past OR would prefer someone Polish (if currently not with anybody)*
   - *0 = non-Polish*

Recent Linguistic Environment

4. Language used at work and/or school
   - *2 = Polish*
   - *1 = mixed*
   - *0 = completely English*
5. Language used at home (with parents or partner)
   - *2 = Polish (or lives with parents)*
   - *1 = mixed (or lives alone)*
   - *0 = English*
6. Language used with friends
   - *2 = Polish*
   - *1 = mixed*
   - *0 = English*

Childhood Linguistic Environment

7. Language used at school
   - *2 = Attended school in Poland*
   - *1 = Attended Saturday heritage language classes*
   - *0 = Attended school in Canada*

8. Language used at home (with parents and/or siblings)
   2 = Polish
   1 = mixed (e.g. English with siblings)
   0 = English
9. Language used with friends
   2 = Polish
   1 = mixed
   0 = English

Language Use

10. How well do you speak Polish?
    2 = good/perfect
    1 = some minor problems/could be better
    0 = major problems
11. How often do you speak Polish?
    2 = everyday
    1 = a few times per week
    0 = once a week
12. Which language do you prefer to speak?
    2 = Polish
    1 = both
    OR tries to speak/practice Polish whenever possible
    0 = English

Cultural awareness

13. Do you watch/listen/read Polish media?
    2 = reads newspapers, watches TV, or listens to the radio often
    1 = will try to sometimes
    OR is interested in some things but not others
    0 = no interest in Polish media
14. Do you visit Poland? How often? For how long?
    2 = lived there for over 3 months since immigrating to Canada
    1 = visit often for at least 1 month at a time
    0 = never or rarely
15. Are you aware of discrimination against Polish people?
    2 = has personally experienced discrimination
    1 = is aware of discrimination and/or stereotypes
    0 = is not aware of any discrimination

APPENDIX F: Answers for revised ethnic orientation measures

The following table summarizes the statements each speaker gave with respect to the revised ethnic orientation measures (see Appendix E). Speaker codes are given in the leftmost column. At the top of each column, each measure is identified by its corresponding number from Appendix E. At the end of the table, the total number of points for each particular speaker is tabulated. This total is then divided by the number of questions (15) to give the average score.

Speaker	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Total	Average
P1M88	2	2	0	0	1	1	2	2	2	2	1	1	2	0	1	19	1.27
P1M60	2	2	2	0	2	2	2	2	2	2	2	2	1	1	0	24	1.60
P1F56	2	2	2	1	2	2	2	2	2	2	2	2	2	1	1	27	1.80
P1M55	1	2	0	0	0	0	2	2	2	2	0	2	1	1	1	16	1.07
P1M44	2	2	0	1	0	1	2	2	2	1	2	1	2	1	2	21	1.40
P1M35	2	2	1	0	1	2	2	2	2	2	1	2	2	1	0	22	1.47
P2M47	1	2	0	0	1	2	0	1	0	0	0	0	0	1	2	10	0.67
P2M29	1	2	0	0	1	0	1	2	1	1	2	0	1	1	2	15	1.00
P2F23	1	1	2	0	2	0	1	1	1	2	2	0	0	2	0	15	1.00
P2M23	1	1	0	1	1	0	0	1	0	1	1	1	1	1	1	11	0.73
P2M22	1	2	1	0	2	1	0	2	0	1	2	1	1	0	1	15	1.00
P2M21	1	2	1	0	2	1	1	1	1	2	2	1	2	1	1	19	1.27
P2F18	1	2	1	1	2	1	1	2	1	2	2	1	1	1	1	20	1.33

APPENDIX G: Language mixing scores

The language mixing score was based on the following 8 measures.

Ethnic Identity

1. Speaker

Recent Linguistic Environment

2. Language used at work and/or school
3. Language used at home (with parents or partner)
4. Language used with friends

Childhood Linguistic Environment

5. Language used at school
6. Language used at home (with parents and/or siblings)
7. Language used with friends

Language Use

8. Which language do you prefer to speak?

Each measure was scored based on the speaker's self-reported usage: no points were given if a mixture of both English and Polish were used; while one point was given if either English of Polish were used exclusively.

Speaker	1	2	3	4	5	6	7	8	Total	Average
P1M88	1	1	0	0	1	1	1	0	5	0.63
P1M60	1	1	1	1	1	1	1	1	8	1.00
P1F57	1	0	1	1	1	1	1	1	7	0.88
P1M55	0	1	1	1	1	1	1	1	7	0.88
P1M44	1	0	1	0	1	1	1	0	5	0.63
P1M36	1	1	0	1	1	1	1	1	7	0.88
P2M47	0	1	0	1	1	0	0	1	4	0.50
P2M29	0	1	0	1	0	1	1	1	5	0.63
P2F23	0	1	1	1	0	1	1	1	6	0.75
P2M23	0	0	0	1	1	0	0	0	2	0.25
P2M22	0	1	1	0	1	1	0	0	4	0.50
P2M21	0	1	1	0	0	0	1	0	3	0.38
P2F18	0	0	1	0	0	1	1	0	3	0.38

APPENDIX H: Linguistic confidence scores

The linguistic confidence score was based on the following two measures. Each of these measures was given one score based on the researcher's (objective) assessment of a speaker, and another score based on the speakers' (subjective) assessment of their language use.

1. a. How well do you speak? (researcher assessment)
   Good = 1
   Bad = 0
   b. Is it good enough? (speaker assessment)
   Yes = +0
   No = +1
2. a. How often do you speak? (researcher assessment)
   Almost everyday = 1
   About once a week = 0
   b. How often do you speak? (speaker assessment)
   Often = +0
   Rarely = +1

The total number of points was divided by the number of questions (2) to determine the average score.

Speaker	1a	1b	2a	2b	TOTAL	AVERAGE
P1M88A	1	0	1	0	2	1.00
P1M60A	1	0	1	0	2	1.00
P1F57A	1	0	1	0	2	1.00
P1M55A	1	0	0	1	2	1.00
P1M44A	1	1	1	0	3	1.50
P1M36A	1	0	1	0	2	1.00
P2M47A	0	0	0	0	0	0.00
P2M29A	1	1	1	0	3	1.50
P2F23A	1	0	1	0	2	1.00
P2M23A	1	1	1	1	4	2.00
P2M22A	1	0	1	0	2	1.00
P2M21A	1	0	1	0	2	1.00
P2F18A	1	0	1	0	2	1.00