Multifactorial Analysis of V1 Constructions in Hebrew Intransitive Clauses

By
Hillel Taub-Tabib

A thesis submitted for the degree of
Master of Arts

Prepared under the guidance of
Prof. Mira Ariel

Department of Linguistics,
The Lester and Sally Entin Faculty of Humanities,
Tel-Aviv University

December 2009

Acknowledgments

First and foremost I would like to thank my supervisor, Prof. Mira Ariel. This thesis could not have been written without her enduring encouragement and support. Her insightful comments have shaped this thesis, and our meetings and discussions have had an invaluable influence on my understanding of language and of linguistic research.

I am also greatly in debt to Prof. Ron Kuzar who provided me with drafts of his ongoing work, and has taken the time to answer all my questions in great detail. This thesis has greatly benefited from his advice.

For helping me with the statistical and computational aspects of this work, I am grateful to Avshalom Koren, Prof. Stefan Gries, Prof. Yoav Benjamini and the participants of the TAU statistics seminar. Their advice and comments have helped me come to terms with many of the issues involved.

I would like to thank Tal Linzen, Dr. Gonen Dori-Hacohen, Prof. Yael Maschler and Prof. Shlomo Izre'el for providing me with corpora for this work. I am also grateful to Yoni Neeman and my colleagues at Melingo Ltd. for providing me with automated software tools to analyze these corpora.

Finally, for their helpful comments and suggestions during the course of this work, I would like to thank Prof. Tal Siloni, Prof. Esther Borochovsky Bar Aba, Dr. Nurit Melnik, Aviva Silbergeld M.Sc (for editing), Roey Gafter and Prof. Ruth Berman and her group.

Contents

Part I

Theoretical Discussion

1 Overview

1.1 The Phenomenon

Hebrew exhibits the unmarked word order SV(O), in where the verb follows its subject. It is however also possible, subject to various constraints, to use the marked word orders VS or VOS in which the subject follows the verb. This phenomenon is often termed *subject-verb inversion* or simply *inversion*. In this thesis I focus on V1 sentences, a subset of inverted sentences where the verb is the first element of its clause[1]. Some V1 sentences are exemplified in (1):

(1) a. acar oti šoter.
arrested me a-policeman.
'A policeman arrested me.'

b. hitxil iti mišehu b-a-mesiba.
flirted with-me someone at-the-party.
'Someone at the party made a pass at me.'

c. nišpax po kafe.
was-spilt here coffee.
'Coffee was spilt here.'

d. nirdam li ha-gav.
fell-asleep to-me the-back.
'My back went numb.'

e. hitxil ha-tekes.
began the-ceremony.
'The ceremony started.'

The sentences in (1) exhibit a wide range of V1 constructions. In (1-a) and (1-b) we see transitive verbs with direct and indirect objects, (1-c) and (1-d) exhibit intransitive verbs with locative and dative modifiers and in (1-e) we see an intransitive verb with no modifiers.

V1 sentences such as the ones in (1) normally have a reasonably acceptable SV alternate with roughly the same meaning (e.g. *šoter acar oti* 'a-policeman arrested me' or *ha-tekes hitxil* 'the-ceremony started')[2]. The inverse however, is not true. Many SV sentences become awkward once their subject is positioned after the verb. The sentences in (2) are syntactically equivalent to the natural sounding inversion examples in (1), but they are nonetheless awkward sounding compared to their SV(O) counterparts in (3).

[1] V1 sentences are also called simple inversion sentences or free inversion sentences. The other type of inverted sentences are triggered inversion sentences where there is a pre-verbal phrase that facilitates the inversion resulting in an [XP V S] word order (e.g. *etmol higi'u ha-orxim* 'yesterday arrived the-guests'). I will discuss the differences between the two types of inverted constructions in section 1.2 and in chapter 2, but I will not discuss triggered inversion in the empirical part of this work.

[2] Exception to this are the existential predicate *yeš 'there-is'* and its inflections *haya* 'there-was' and *yihiye* 'there-will-be' whose word order is for the most part fixed to verb–subject (except in rare cases of contrastive focus or in archaic literary use).

(2) a. ? bana et ha-ca'acu'a yeled.
built ACC the-toy a-child.
'A child built the toy.'

b. ? hitxila im mišehu dana b-a-mesiba.
flirted with someone dana at-the-party.
'Dana made a pass at someone at the party.'

c. ? kafcu al-ha-šulxan yeladim.
jumped on-the-table children.
'children jumped on the table'

d. ? na'am ha-politika'i.
gave-a-speech the-politician.
'The politician gave a speech.'

(3) a. yeled bana et ha-ca'acu'a.
a-child built ACC the-toy.
'A child built the toy.'

b. dana hitxila im mišehu b-a-mesiba.
dana flirted with someone at-the-party.
'Dana made a pass at someone at the party.'

c. yeladim kafcu al-ha-šulxan.
children jumped on-the-table.
'children jumped on the table'

d. ha-politika'i na'am.
the-politician gave-a-speech.
'The politician gave a speech.'

So what is it that motivates the choice of VS order in (1) and at the same time renders it awkward in (2)? Over the years, qualitative research has revealed many factors that are argued to correlate with inverted orders, possibly motivating their choice (cf. Givón, 1976a, Shlonsky, 1987, 1997, Kuzar, 1990, 2006b,a, forthcoming, Melnik, 2002, 2006). The proposed factors span different domains of linguistic analysis. Morpho-syntactic factors such as definiteness and NP type are considered, along side semantic factors like animacy and unaccusativity and along side pragmatic ones such as accessibility and topicality[3]. The multiplicity and co-dependency of the factors poses a problem for our understanding of the inverted word order's motivation—the factors are strongly dependent and correlated so it is hard to argue which of them (if any) motivates word order and which are epiphenomenal. For instance, inanimate subjects are much more frequent in V1 than in S1 sentences, so one can argue that animacy (at least in part) lies behind the choice of an inverted word order. However, it is well known that inanimate subjects frequently coincide with non-agentive verbs, and indeed, data shows that the vast majority of V1 sentences with inanimate subjects also have non-agentive verbs. It is thus difficult to decide if it is non-animacy or non-agentivity that contribute to the choice of the V1 word order. It is also possible that both factors contribute to this choice, or that both factors are epiphenomenal to a third factor. These questions are best answered using quantitative methods and they will be the subject of chapters 4 and 5 of this work.

[3] Definiteness and NP type are features relating to the subject argument, where it is suggested that inverted sentences prefer indefinite and lexical subject over definite and pronominal ones. Animacy relates to the subject where it is suggested that inverted sentences prefer inanimate over animate subjects. Unaccusativity relates to the verb's semantic class where it is suggested that inverted sentences prefer unaccusative verbs over unergative ones. Accessibility and topicality again refer to the subject argument, but they are discourse related factors that correlate to the degree of salience the subject referent has in the hearer's mind (accessibility), and the degree to which the subject is interpreted as "what the sentence is about" (topicality). All these factors will be discussed in detail in subsequent chapters.

When discussing motivations to the V1 word order, it is important to consider two separate questions: (i) which factors diachronically motivate the availability of different word orders and their association with different linguistic features, and (ii) which factors affect the synchronic choice of word order in a given discourse context (for lack of a better term, I will refer to this choice as the "online" choice of word order). My answer to the first, diachronic question, largely following Givón (1976a) and Lambrecht (1994, 2000), is that the availability of different word orders, as well as their association with different linguistic features, can be accounted for by a single motivation—the need to mark non-topical subjects. As for the second question, I will differ from current approaches by introducing a probabilistic multifactorial account. I will argue that no one factor can account for the "online" choice of word order, and that the problem is best modeled by the use of multiple factors (cf. Gries, 2003, Bresnan et al., 2007).

Lambrecht (1994) argued that grammatical forms (such as word orders or intonation patterns) arise diachronically "under pressure" of information structure considerations. He also demonstrated that these information structure considerations can account for the alignment of the grammatical forms with specific values of different linguistic features (e.g. the alignment of VS order with non-agentive verbs, indefinite/non-human subjects etc.). Continuing this line of thought, I will argue that the human mind picks up on these correlations and proceeds to softly associate these feature values with the respective word orders. After a while, this process results in a situation where a given grammatical form (or in our case—word order) is no longer associated only with the information structure configuration that diachronically motivated it, but rather with a set of features that motivate its "online" choice . The different features can of course affect the choice of word order to different degrees. The assignment of precise weights to the different features is best achieved through a quantitative examination and is the subject of the second part of this research.

This thesis is comprised of two parts: a theoretical discussion (chapter 2-3) and an empirical investigation (chapters 4-5). It is outlined as follows: In chapter 2 I introduce current approaches to V1 constructions while highlighting their pros and cons. In chapter 3 I introduce the idea that the motivation for the availability of the VS word order and for its alignment with different feature values is the need to mark non-topical subjects. I end that chapter by showing that synchronically, the need to mark non-topical subjects is not enough to single-handedly account for the "online" choice of word order. In chapter 4, the chapter opening the empirical part of the work, I discuss the methodology of my corpus based empirical analysis and introduce the different features I consider and the methods I use to deduce their values from corpus data. In chapter (5) I detail the results of the quantitative study, arrive at a set of factors that best account for the range of data, and present a working computational model for the prediction of word order.

1.2 The Scope of This Study

Given the title of this work "Multifactorial analysis of V1 constructions in Hebrew intransitive clauses" three terms require clarification: (i) V1 constructions, (ii) Hebrew, and (iii) intransitive clauses.

V1 constructions, or when instantiated, V1 sentences, are sentences in which a verb occupies the sentence initial position, and is subsequently followed by a phonologically expressed subject argument (there may be additional modifiers between the verb and its subject). This definition excludes null-subject sentences, and importantly, it excludes sentences of the type in (4-a), which are called V2 sentences or triggered inversion sentences (because of the pre-verbal inversion trigger). Triggered

inversion sentences (hereafter TI sentences) are more accommodating to different types of verbs than V1 sentences, and they are appropriate in a wider range of situation types. One can add a pre-verbal trigger to practically every V1 sentence, but the inverse does not hold. Many triggered inversion sentences will sound awkward without their trigger.

(4) a. lefeta kafca me-ha-sixim arnevet levana.
suddenly jumped from-the-bushes a-hare white.
'A white hare suddenly jumped from the bushes.'

b. ? kafca me-ha-sixim arnevet levana.
jumped from-the-bushes a-hare white.
'A white hare jumped from the bushes.'

(5) a. nigmar ha-xofeš.
ended the-holiday.
'The holiday ended.'

b. etmol nigmar ha-xofeš.
yesterday ended the-holiday.
'The holiday ended yesterday.'

The sentence pair in (4) exhibits a triggered inversion sentence in (4-a) and its less acceptable V1 counterpart in (4-b). The sentence pair in (5) exhibits a V1 sentence in (5-a) and its triggered inversion counterpart (5-b). In this case the TI counterpart sounds just as natural. These data raise a question concerning the classification of (5-b). Since this sentence is acceptable with and without the trigger, can it perhaps belong to the V1 class rather then the triggered inversion class? Can we treat it as a V1 sentence that just happens to have a prefixed temporal adjunct? In this work I will avoid this characterization, mainly because of the empirical difficulties it introduces[4]. I will will abide by the definition I provided above and consider only sentences that open with a verb.

As for Hebrew, I refer to the dialect which Hebrew speakers employ in their everyday conversations. This definition excludes literary registers, but it does not necessarily exclude all written texts. Specifically, the corpus used for the empirical part of this work is Linzen's blogs corpus (Linzen, 2009). It contains texts from various bloggers that discuss their day to day activities, for the most part using their everyday Hebrew.

Finally, intransitive clauses are clauses in which the predicate is not accompanied by a direct or indirect object. In the part I of this work (the theoretical discussion) I do discuss some transitive examples, but my corpus data is too sparse to allow quantitative analysis of these constructions.[5].

[4] The empirical part of this work relies on corpus analysis rather than on sentence judgments. When I encounter an [XP V S] sentence in the corpus, it is impossible for me to decide without resorting to judgments if the sentence will sound natural in the [V S] order as well.

[5] The V1 word order is much more common in intransitive clauses than it is in transitive ones (my V1 corpus sampled from Linzen's blogs corpus contained 17 transitive V1 sentences vs. 370 intransitive V1 sentences). This phenomenon was shown by Sornicola (2006) to be cross linguistic. Sornicola (2006, p. 456) also reports the results of Uhlířová (1969) which statistically demonstrate (for Czech texts) that increasing the number of sentence constituents increases the rigidity of word order. Apparently the vast number of alternative orderings in sentences with multiple constituents paradoxically urges the speakers to constrain their choice of word order. This phenomenon is robust but not yet fully understood.

2 Existing Approaches

2.1 Syntactic Unaccusativity

2.1.1 Introduction

In an attempt to outline a syntactic account of the Hebrew clause structure, Shlonsky (1997) emphasized the role of the verb's argument structure in determining word order. He examined both triggered inversion, and V1 constructions (in his terminology, free inversion, hereafter FI), and argued that while TI is oblivious to the type of the verb and allows for definite subjects, FI is only possible with verbs whose subject is an internal argument (unaccusatives and passives)[6], and only with indefinite subjects.

In the following sections Shlonsky's theoretical assumptions will be discussed (section 2.1.2), as well as his account for free inversion (section 2.1.3). In section 2.1.4 I will criticize his account by showing that some of its basic assumptions are falsified by corpus data. Shlonsky's account for triggered inversion is not directly relevant to this thesis, but for completeness it is concisely reviewed and criticized in Appendix A.

2.1.2 Syntactic Assumptions

Shlonsky grounds his work in current generative syntactic theory. He views clauses as composed of three layers: (i) the thematic layer, the VP, comprised of the predicate and its θ-marked complements; (ii) The functional layer, IP, comprised of functional projections such as Asp(ect)P, T(ense)P, etc.; and (iii) the operator layer, CP, comprised of Comp and related projections. Figure 1 diagrams this hierarchy.

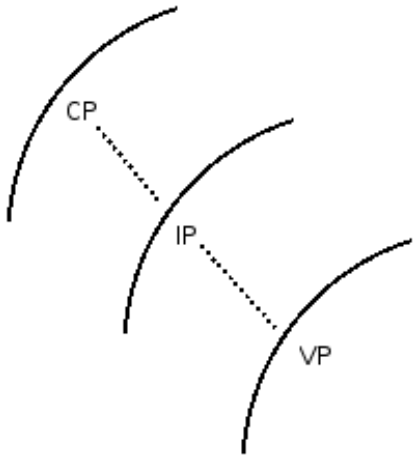

Figure 1: A schematic view of Shlonsky's clausal hierarchy.

Shlonsky does not specify the internal structure of the CP layer. His hierarchy for functional heads comprising the IP is specified in (6) below:

[6] The unaccusative verb class contains mainly verbs of existence, appearance and externally caused change of state . In generative grammar it is assumed that this class of intransitive verbs is differentiated from another class, the class of unergative verbs (mainly agentive verbs and verbs of internal causation such as *sparkle or blossom*), in that the subject of the unaccusative verbs is an internal argument (positioned within the VP in D-structure. An underlying object) while the subject of unergatives is an external argument. For an overview of the unaccusative and unergative verb classes see Levin and Rappaport Hovav (1995).

(6) The IP Hierarchy

a. Simple Tenses
AgrSP > (NegP) > TP > AspP > AgrOP > VP

b. Compound Tenses (Benoni)
AgrSP > (NegP) > TP > VP> AgrPartP > AspP > AgrOP > VP

As for the VP, Shlonsky adopts the VP internal subject hypothesis given in (7):

(7) The subject is internal to the VP and originates in [spec, VP].

Finally, Shlonsky (1997, p. 71) argues for verb movement of the lexical verb outside the VP boundaries. Using adverb placement diagnostics, he demonstrates that the verb is not within the VP in surface structure and concludes that movement must have taken place. Regarding the exact position of the verb within the IP he finally concludes:

(8) Finite (past/future) and non-finite verb raise to $AgrS^0$ while benoni/present verbs raise to T^0.

Wherever possible, I will abstract from (8) and refer to the verb's position as simply [spec, IP].

2.1.3 Free Inversion

In contrast to TI sentences, where inversion is always possible due to the presence of a trigger, FI sentences, Shlonsky argues, are only appropriate with verbs that have their subject as their internal argument (unaccusatives and passives) and only with indefinite subjects. These assumptions are imperative to his syntactic analysis of the phenomenon, and he demonstrates them using the acceptability judgments in (9) repeated from (Shlonsky, 1997, p. 163)[7].

(9) a. ne'elmu harbe sfarim me-ha-sifriya.
disappeared many books from-the-library.
'Many books disappeared from the library.'

b. * ne'elmu ha-sfarim me-ha-sifriya.
disappeared the-books from-the-library.
'The books disappeared from the library.'

c. be-šavu'a šeavar ne'elmu ha-sfarim me-ha-sifriya.
on-the-week the-last disappeared the-books from-the-library
'Last week, the books disappeared from the library.'

ne'elam 'disappeared' is an unaccusative verb, and is thus acceptable in the FI sentence in (9-a), but such a sentence is only possible insofar as the subject is indefinite, which accounts for the acceptability difference between (9-a) and (9-b). (9-c) demonstrates that in TI sentences both definite and indefinite subjects are possible.

[7] Shlonsky's acceptability judgments about the definiteness effect in FI sentences have been contested before by Melnik, who argued, based on her own and her informants' judgments, that sentences Shlonsky marks as unacceptable are in fact quite acceptable. While I agree with this sentiment, for now I cite Shlonsky's arguments as is. I will critique them using corpus evidence in the next section (2.1.4).

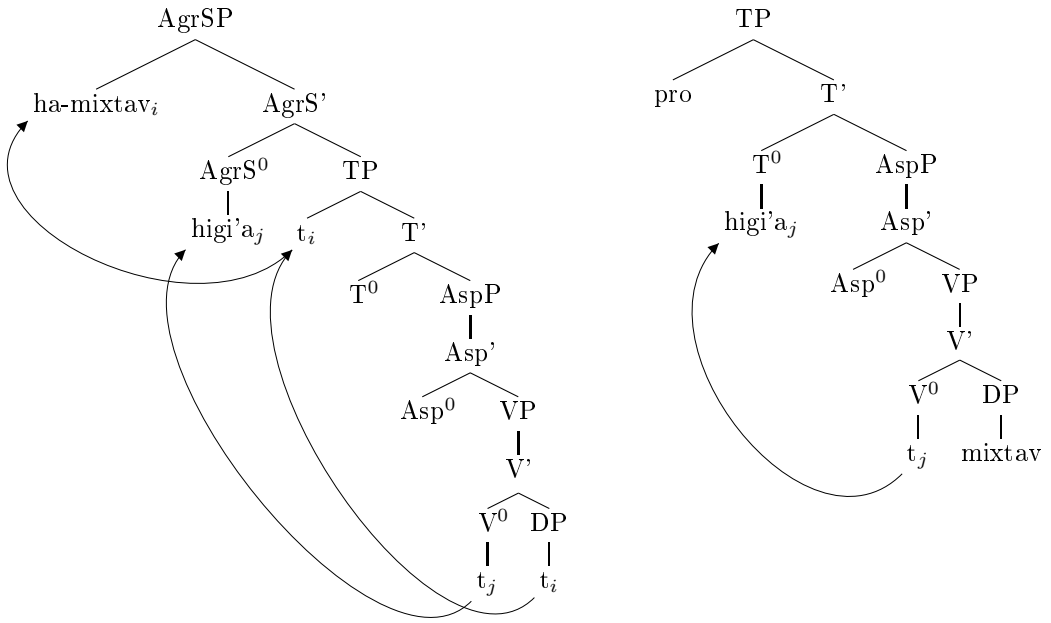

Figure 2: Shlonsky's detailed analysis for the SV and FI sentences

Shlonsky's account for these data is syntactic. He argues that subjects of transitives and unergatives cannot remain in the VP because they must be case licensed by I_0 while the subjects of unaccusative and passive verbs can be licensed by V_0 directly and can thus stay within the VP, which results in a verb-subject order. He assumes along with Belletti (1988) that passives and unaccusatives can license their subject since they can assign partitive case, as opposed to unergatives and transitives, which only assign accusative. It is known that partitive case can only be assigned to indefinite subjects and this seems to correlate perfectly with the putative ban on definite subjects in FI sentences. Finally, it should be noted, that in FI sentences the EPP is maintained by *pro* which occupies the [spec, IP] position[8]. In fact, it is argued by Reinhart and Siloni (2004a, footnote 10) that it is the selection of *pro* as a lexical item for the derivation that determines the word order of the sentence: if *pro* is selected then there is no need for the subject to move in order to maintain the EPP resulting in the verb-subject word order. Otherwise, if *pro* is not selected, the subject will have to move to [spec, IP] and will thus move past the I^0 verb resulting in a subject-verb order. Figure 2 details Shlonsky's analysis for the SV sentence *ha-mixtav higia* 'the-letter arrived' along with its FI counterpart *higi'a mixtav* 'arrived a-letter'.

2.1.4 Criticism

Shlonsky's account of free inversion hinges on two assumptions which appear to be contradicted by corpus evidence:

(10) No unergative or transitive verbs can appear in FI sentences.

[8] The EPP is a principle of generative linguistics according to which clauses must have subjects (where subjects are taken to be elements in [spec, IP]). In null subject languages such as Hebrew, the EPP requirement may be maintained by the phonologically null element *pro* (cf. Chomsky, 1981).

(11) The subject of FI sentences must be indifinite.

Both assumptions are crucial to Shlonsky's analysis of FI. Assumption (10) is essential because the only way in which an unergative verb can precede its subject, is if the subject remained in the VP (see (8) above and bear in mind that subjects that move outside the VP occupy [spec, AgrSP]) but that would contradict his claim that subject of unergatives cannot be assigned case there. Assumption (11) is also critical because Shlonsky uses it in order to explain how unaccusatives assign case to their VP internal subjects.

I will address these two assumptions in order. As counterexamples to the first assumption we can observe the examples in (12), all attested examples from Linzen's blogs corpus:

(12) a. yašva tamar ve-xikta be-beyt aviha.
sat tamar and-waiter in-house(GEN) her-father.
'Tamar sat and waited in her father's house.'

b. yašen eclenu ha-ben šel gisi.
sleeps at-us the-son of my-brother-in-law.
'My brother in law's son is sleeping at our house.'

c. anta mišehi še-amra še-hi mekabelet hoda'ot avur šerut ha-lakoxot.
Answered someone that-said that-she accepts messages for the customer service.
'Someone answered and said that she takes messages for the customer service.'

d. azvu otanu kama anašim yekarim.
left us a-few people dear.
'A number of beloved people have left us.'

e. patxa l-i et ha-delet ozeret-bayit rusiya ve-xaviva.
opened to-me ACC the-door cleaning-lady Russian and-friendly.
'A friendly Russian cleaning lady opened the door for me.'

The verbs *yašav* 'sat', *ya*šan 'slept' and *ana* 'answered' exemplified above, are all unergative verbs and they are all frequent in FI sentences. In section 2.3.3 I will argue that while unaccusative are far more frequent than unergatives in FI sentences, this is only a tendency. Given an appropriate context almost any verb can appear in FI sentences[9]. The verbs *azav* 'left' and *patax* 'opened' in (12-d) and (12-e) are transitives, again, contra to Shlonsky's assumption in (10).

Counterexamples to the second assumption are even easier to come by. Indeed, about a full half of the FI sentences in my sample of the Linzen corpus had definite subjects. Some examples are presented in (13), many more were attested:

[9] Alexiadou (2007) based on Borer (2005) criticized the use of free inversion as an unaccusative diagnostic, among other reasons, because of the fact that unergatives sometimes appear in V1 constructions when there is an intervening locative between the verb and its subject (e.g. example (12-b)) . It was pointed out to me by Tal Siloni (personal communication) that a categorical syntactic constraint against unergative verbs in V1 constructions might still be viable if we limit our discussion to strict [V S] sentences (verb initial sentences where there is no intervening modifier between the verb and its subject). However, as can be seen in examples (12-a) and (12-c), unergative verbs can appear in strict [V S] sentences even without such modifiers. Siloni points out that perhaps if we limit discourse context to "out of the blue" sentences, then the constraint holds. However, I am not certain of that either as it seems to me that the following exchange is felicitous: Q: *xazarti! ma kore?* 'I'm-back! what's going on?' A: *cilcel moše me-hamakolet ve-amar ...* 'called Moshe from-the-grocery-store and said ...'. Despite all this, it is definitely the case that unergatives in V1 constructions are much more frequent when there is an intervening modifier. If one wishes to argue for a syntactic constraint they will have to accommodate the above counterexamples, and perhaps more importantly, provide a syntactic account for the fact that V1 unergatives do appear in the presence of intervening modifiers (and also occasionally in strict [V S] sentences)—all these steps are absent from Shlonsky's account.

(13) a. hitxil ha-tekes.
began the-ceremony.
'The ceremony began'

b. nišbera li ha-kos.
broke to-me the-glass.
'The glass was broken.'

To summarize, Shlonsky's syntactic account for free inversion is based on constraints against the appearance of unergative/transitive predicates and definite subjects in FI sentences. These constraints appear to be empirically untenable. In order to account for the full range of data we will have to consider other factors beyond the verb's argument structure.

2.2 V1 Sentences as Thetic Sentences

2.2.1 Introduction

Melnik (2002, 2006) motivates the choice of the V1 word order in terms of the distinction between thetic and categorical judgments (or propositions). Categorical judgments are propositions that consist of two acts: the act of naming an entity and the act of making a statement about it. Thetic judgments on the other hand, are viewed as a logically simple expression of an event or situation. Melnik argues that V1 constructions are the mechanism used in Hebrew to express thetic judgments. Accordingly, the function of the inverted word order is to differentiate thetic judgments from categorical ones.

In what follows I will review a number of approaches to theticity, and then discuss Melnik's approach and its shortcomings. I will argue that in order to use the term in a way that aligns best with Hebrew V1 constructions, one has to adopt an interpretation that regards thetic sentences as sentences whose subject is not topical, effectively rendering it equivalent to the proposal I present in this thesis. I will conclude by arguing that such an interpretation of theticity is not only consistent with Hebrew V1 constructions, but is also the most effective one cross-linguistically.

2.2.2 Thetic and Categorical Judgments

The distinction between thetic and categorical judgments originates in the theories of Brentano (1874) and Marty (1918), and was adapted to modern linguistics by Kuroda (1972). The term judgment was used in the early works and it relates to the way the speaker perceives the situation she is reporting on. Kuroda (1972, p. 154) argued that it might be appropriate to replace it with a more modern term such as proposition or statement, but he retained it in his 1972 paper for convenience (in order to remain consistent with Brentano and Marty's terminology). Thus terms such as thetic judgments and thetic propositions are used somewhat interchangeably today. Kuroda explained the distinction in the following passage (Kuroda, 1972, p. 154):

> "This theory assumes, unlike either traditional or modern logic, that there are two different fundamental types of judgments, the categorical and the thetic. Of these, only the former conforms to the traditional paradigm of subject–predicate, while the latter represents simply the recognition or rejection of material of a judgment. Moreover, the categorical judgment is assumed to consist of two separate acts, one the act of recognition of that

> which is to be made the subject, and the other, the act of affirming or denying what is expressed by the predicate about the subject. With this analysis in mind, the thetic and the categorical judgments are also called the simple and the double judgments.

It is important to note that when Kuroda mentions subject–predicate he does not refer to the grammatical subject in the sense I will be using throughout this thesis (see section 3.2 and Appendix B), but rather to the element that the sentence is about, i.e. the topic[10]. In current terms, it can be said that the categorical judgment conforms to a topic–comment paradigm, while the thetic judgment is topicless.

Kuroda used the thetic/categorical distinction to account for particle selection in Japanese (*ga/wa*). He exemplified the distinction with the following sentence pair:

(14) a. Inu ga hasitte iru.
a/the dog PAR is running.
'A/The dog is running.'

b. Inu wa hasitte iru.
the dog PAR is running.
'The dog is running.'

Kuroda explains that in a situation where an English speaker notices a dog running in the street and says *a dog is running*, a Japanese speaker would use the sentence with the particle *ga*. The reason is that the speaker perceives or judges the situation he wishes to report as event central. His goal is not to convey some new information about the dog, but rather to report an event of running in which the dog happens to participate. Kuroda suggests to analyze such judgment as:

(15) a. Running of X.
b. X is a dog.

Kuroda later emphasizes that a situation can be judged as thetic even in cases where its participant is discourse old:

> Consider the same situation in which a dog is running ... but assume that the dog is not an arbitrary dog but a certain definite dog familiar to the speaker or whose identity has already been established to the speaker and hearer. As in the previous case, the speaker recognized X's running ... but the speaker refers to X perhaps by the dog's name, say, *Fido*, in case the name is known to him, or perhaps by some definite noun phrase like *the dog* in case the identity of the dog has been otherwise established.

(16) a. Fido ga hasitte iru.
Fido PAR is running.
'Fido is running.'

[10] In his paper, Kuroda writes that his concept of subject should be distinct from the 'topic', but that is only because he considers topic to be 'old information'. He is actually arguing that the concept of 'what the sentence is about' should be separate from the old/new information dichotomy. I accept this point and discuss it in section C.3. However, since my definition of topic is in terms of aboutness alone, it results that Kuroda's subject is exactly what I am calling topic.

b. Inu ga hasitte iru.
The dog is running.
'The dog is running.'

In Kuroda's terms a sentence like *Fido is running* is topicless in cases where the speaker perceives it as an event reporting sentence that just happens to involve Fido. We can understand then, that Kuroda's definition of topic is cognitive, it refers to an element about which the speaker intends to add information, and not just an element about which the speaker happens to add knowledge because it is participating in an event the speaker is reporting.

Kuroda further notes, that if a sentence has any topical elements it should be considered categorical. He presents the following examples:

(17) a. Inu wa niwa de neko o oikakete iru.
the-dog PAR in the garden cat PAR is chasing.
'The dog is chasing the cat in the garden.'

b. Neko wa inu ga niwa de oikakete iru.
the-cat PAR a/the-dog PAR in the garden is chasing.
'The cat is being chased by a/the dog in the garden.'

c. Niwa de wa inu ga neko o oikakete iru.
in the garden PAR a/the-dog PAR cat PAR is chasing.
'In the garden, a/the dog is chasing a/the cat.'

Kuroda argued that Japanese reflects the thetic/categorical distinction through its *wa*/*ga* marking. He made no attempt to argue that other languages reflect it as well, and in fact, he implicitly argued that English does not reflect it by arguing that in English a sentence such as *The dog is running* is ambiguous between a thetic and categorical reading.

Melnik (2002, 2006) defines thetic sentences in a way similar to Kuroda's. In Melnik (2002, p. 159) she writes:

> The distinction between thetic and categorical expressions, then, is that categorical expressions are 'about something' while thetic expressions are not. Thus, categorical expressions contain a 'predication base' while thetic expressions do not.

While Melnik does not explicitly call thetic sentences topicless, her definition in terms of aboutness and predication base appears to be equivalent. Aboutness and predication base are in themselves terms used to define topics and their lack—a way to define topicless sentences. This definition leads Melnik to introduce a caveat to the generalization that V1 constructions are used to encode thetic expressions. The caveat relates to the [V O S] and [V DAT S] constructions in (18):

(18) Q: What happened?/What happened to you?

a. aktsa oti dvora.
Stung me a bee.
'A bee stung me.'

b. nikre'u li ha-mixnasayim.
tore to-me the-pants.
'My pants tore.'

Melnik states that when the context is the second question (*what happened to you?*) there is no way to argue that the answer is not 'about something' and thus the sentences are categorical. She concludes that the [V O S] and [V DAT S] constructions are ambiguous between a thetic expression and categorical one, and that in their categorical guise the predication base is the O/DAT argument. I would add, that in the examples above it does not really matter if the question is *What happened?* or *What happened to you?* Whenever a speaker asks an addressee *What happened?* and where the answers above are felicitous, I believe it is contextually likely that the question regards the addressee.

With this issue in mind, Lambrecht (1994) devised a more inclusive definition of theticity. Observe the following examples from (Lambrecht, 1994, p. 137):

(19) Q: How's your neck?

a. My neck HURTS.

b. Il collo mi fa male. (Italian)
the neck me hurts.

c. Mon cou il me fait mal. (French)
my neck it me hurts.

d. Kubi wa itai. (Japanese)
Neck PAR hurts.

(20) Q: What's the matter?

a. My NECK hurts.

b. Mi fa male il collo. (Italian)
me hurts the neck.

c. J'ai mon cou qui me fait mal. (French)
I my neck have me hurts.

d. Kubi ga itai. (Japanese)
Neck PAR hurts.

Lambrecht referred to the sentences in (19) and (20) as allosentences and described them as semantically identical but pragmatically distinct. Their pragmatic function, according to Lambrecht, is to encode the thetic and categorical expressions in the different languages. English contrasts accented and non-accented subjects, Italian contrasts post verbal and preverbal subjects, French contrasts clefted and detached subjects and Japanese marks the subjects (ga vs. wa). According to Lambrecht (2000) the manifestation of the thetic category ('Sentence Focus' in his terminology) is motivated by a single principle - the principle of paradigmatic contrast, that is, the need to be minimally distinct from the corresponding categorical ('Predicate Focus') structure. Lambrecht claims that this is achieved by *detopicalization* of what is prototypically the topic. In the process of detopicalization, the subject loses some of its subject properties in a process of *subject-object neutralization*.

It is already evident from the examples above that Lambrecht's concept of theticity is distinct from that of Kuroda and Melnik's. Examining (20-a) it is safe to assume that the speaker intends to convey information about himself, and thus the pronoun *my* represents the speaker as the topic. This sentence would be considered categorical by Kuroda and Melnik, but it is one of Lambrecht's favorite examples for a thetic sentence and is now commonly discussed in other papers on theticity as well (cf. Sasse, 2006). Lambrecht (1994, p. 144,145) explains:

> I would like to emphasize that the formal contrast between the marked category of thetic

> sentences and the unmarked category of topic-comment (or categorical) sentences crucially involves the grammatical relation SUBJECT [...] It is not the absence of any topic relation that characterizes thetic sentences but the absence of a topic relation between the proposition and that argument which functions as the topic in the categorical counterpart [...] in the unmarked case this categorical topic argument is the subject. It is in principle possible for non-subject constituents to have topic status in thetic sentences [...] What counts for the definition of the formal category "thetic sentence" is that the constituent which would appear as the subject NP in a corresponding categorical allosentence gets formally marked as NON-TOPIC, resulting in a departure from the unmarked pragmatic articulation in which the subject is the topic and the predicate the comment.

Lambrecht's idea of theticity is then different from Melnik and Kuroda's in that he does not consider thetic sentences to be topicless, but rather sentences in which the subject is not the topic. The two approaches to theticity coincide with regard to canonical (topicless) thetic sentences such as *it's raining* or *there is a god,* but Lambrecht's approach allows for the inclusion of many sentence structures which include a topical element that is not the subject. These structures prove quite prevalent in Hebrew V1 sentences, hence the advantage of this approach to theticity with regard to the phenomenon at hand.

2.2.3 Criticism

My criticism of Melnik's proposal has been implicitly stated in section 2.2.2. Melnik's definition of theticity appears to address canonical thetic sentences such as *yored geshem* 'it's raining' or *yeš elohim* 'There is a God', and other [V S] sentences such as *hitxil ha-tekes* 'began the ceremony' or *nigmar merc* 'March ended'. But it excludes many [V O S] and [V DAT S] sentences and thus excludes a large portion of V1 sentences. The sentences in (21) are just a few of the examples from my sample of the Linzen corpus:

(21) a. nigmar l-i ha-xofeš.
ended to-me the-vacation.
'My vacation ended.'

b. hitxil iti mišehu ben esrim ve-štayim.
flirted with-me someone of-age twenty and-two.
'Some twenty two year old man made a pass at me.'

c. hištatku li ha-raglayim.
became-silent to-me the-feet.
'My feet went numb.'

In fact 37.5% of the V1 examples in my sample of Linzen's corpus were of a [V O S]/[V DAT S] structure, and in most cases the O/DAT was the topical element. Adopting Melnik's definition will exclude these sentences for no good reason. Lambrecht's definition, on the other hand, is equivalent to the approach discussed in chapter 3, and results in a better empirical coverage of the data.

2.3 P1 Situation Types

2.3.1 Introduction

Kuzar (1990, 2006b,a, forthcoming)[11] argues that the choice of word order is determined by a mapping between propositions expressing situation types and sentence patterns which in turn determine their form. This mapping relies on a combination of semantic and pragmatic considerations and the concept as a whole is similar in many ways to other works in construction grammar (cf. Goldberg, 1995, 2006)[12]. In what follows I will introduce the details of sentence patterns and their semantic organization (sections 2.3.2 and 2.3.3), and go on to discuss and criticize some aspects of this approach relating to the role of information structure in motivating word order (section 2.3.4).

2.3.2 Sentence Patterns

Kuzar (forthcoming) compares a sentence pattern to a multi-dimensional cube, whose dimensions are: mood, polarity, modality, information structure and word order. Once a proposition describing a certain situation type is associated with a sentence pattern, the pattern will take into account all of the proposition's parameters and yield its grammatical form (in our case, its word order). For the Hebrew patterns discussed in Kuzar (forthcoming), the word order within a pattern is for the most part fixed, so mapping a proposition to a sentence pattern will effectively determine its word order. The mapping itself is done by considering the semantics and information structure properties of the proposition and matching it with the available options offered by the patterns available in Hebrew. Semantically, the sentence patterns are organized in a prototype structure so it is possible for a proposition to fit more than one pattern.

Kuzar's sentence patterns for Hebrew can be broadly split into two types, the S1 (subject first) sentence patterns and the P1 (predicate first) sentence patterns. The S1 sentence patterns include the verbal sentence pattern, *V S-pattern*[13], which is the home of volitional actions (*dan axal tapu'ax* 'Dan ate an apple'), and the copula sentence pattern, *COP S-pattern*, which provides background information about discourse entities (*dan adam tov* 'Dan is a good person'). The P1 sentence patterns are further divided into major and minor sentence patterns. The major sentence patterns include the existence sentence pattern, *EX S-pattern*, the evaluative sentence pattern, *EV S-pattern*, and the sentence pattern of environmental conditions, *ENV S-pattern*. The minor sentence patterns are used with deteriorating entities, body part conditions, animal induced conditions and cost expressions. Table 1 lists situation types that are expressed in P1 sentence patterns, along with example sentences[14].

[11] Kuzar was kind enough to provide me with his yet unpublished book about sentence patterns (Kuzar, forthcoming). In the course of this thesis I have reviewed and cited different drafts of this book. I have made an effort to update the page numbers and citations so as to fit the book's final draft but obviously discrepancies may exist between the information provided here and the book's published version.

[12] There are however differences between Kuzar's approach and construction grammar, especially concerning the details of the two formalisms. The reader is referred to Kuzar (forthcoming, chapter 1) for an overview of the two formalisms and their differences.

[13] Kuzar uses the term V S-pattern (i.e. the verbal sentence pattern) to describe the canonical S1 sentence. This turns out to be a bit confusing in the context of this work since I often use the term VS sentences (i.e. verb-subject sentences) for the exact opposite. In order to avoid confusion, when using Kuzar's term the exact notation above will always be used (i.e. V space S dash pattern) .

[14] Kuzar designated a sentence pattern for animal induced conditions. In table 1 I have taken the liberty to rename it to the transitive object sentence pattern. Kuzar considered examples such as *akca oti dvora* 'stung me a-bee', but the same construction is used for other situation types that involve direct or indirect objects in which the subject is not the topic (e.g. *acar oti šoter* 'arrested me a-policeman' or *hitxil iti mišehu b-a-mesiba* 'flirted with-me someone at-the-party', see also the attested examples from the Linzen corpus in (12)).

The sentence patterns form a field, which is diagrammed in Figure 3 (cf. Kuzar, forthcoming, p. 162).

To summarize, Kuzar argues that when faced with the need to report a particular situation type, the speaker selects a sentence pattern appropriate for the situation, and the sentence pattern determines the word order. Sometimes a situation has a meaning or pragmatic structure that are closely related to those of more than one sentence pattern. In these cases, a situation can be mapped to more than one pattern. The mapping process is influenced mostly by the semantics and pragmatics of the proposition describing the situation, and it will be considered in the next two chapters.

2.3.3 Conceptual Categories and the Existential Construction

As previously discussed in section 2.1.3, Shlonsky (1997) argued that action verbs (which are unergatives) are impossible in V1 sentences, and that only unaccusative verbs are allowed there. This observation can be accounted for by Kuzar's theory as well, if we notice that the V S-pattern is the home for actions, and V1 S-patterns are associated with situations of existence and appearance, which are normally denoted by unaccusative verbs. However, it was shown in example (12) that agentive actions do infrequently appear in P1 sentences and we argued that this presents a problem for a theory of syntactic constraints such as Shlonsky's.

Observe another example from the daily newspaper Ha'aretz. The speaker quoted, is a person who was attacked by a group of boys[15]:

(22) a. kama dakot axar-kax halaxti leyad migraš ha-kaduregel. racu le-kivuni šlošet
few minutes later I-went near court the-soccer. ran in-my-direction three
ha-ce'irim.
the-young.
'A few minutes later I walked by the soccer court. The three children ran towards me.'

The event denoted by (22) is an agentive event of running. How can its meaning fit that of any of Kuzar's P1 sentence patterns?

To answer this, we should note that Kuzar defines the meaning of his major sentence patterns in terms of conceptual categories. The internal organization of a conceptual category (hereafter CC) is that of a prototype based radial category in which the core is unmarked and the periphery becomes progressively more marked (cf. Lakoff, 1987, p. 91-117 and Kuzar, forthcoming, p. 118-120). It can be thought of as a series of rings, with the core meaning at the inner ring, and the periphery progressively enclosing it. The relation between the core and the periphery is such that the unmarked core meaning is always implied by the more specific peripheral meaning.

A CC is organized pragmatically as well as semantically. Thus, in the core meaning and its close rings we find predicates with a strong lexical-semantic meaning corresponding to the CC, and in the further away rings, we'll see predicates that do not inherently carry the CC meaning, but can attain the meaning through the combination of the discourse context, the meaning embedded in the sentence pattern, and the occasional presence of various modifiers.

From the three P1 conceptual categories surveyed in Kuzar (forthcoming), the CC of existence is the one of most interest to us. It covers the vast majority of V1 sentences (in both token and type, see

[15] http://www.haaretz.co.il/hasite/pages/ShArtPE.jhtml?itemNo=503255&contrassID=2&subContrassID=21&sbSubContrassID=0

Table 1: P1 Situation Types

Sit. Type	Example
Existence/Negative Existence	yeš bxirot 'EXIST elections'
Presentation	higi'a rakevet, 'arrived a-train'
Possession	yes l-i sefer 'EXIST to-me a-book'
Negative Possession	ein l-i sefer 'NEG EXIST to-me a-book', ne'elam l-i ha-darkon 'is gone to-me the-passport'
Deteriorating Entity	nikre'a l-i ha-xulca 'was-torn to-me the-shirt', hitkavcu ha-mixnasayim 'shrank the-pants'
Body-part condition	nishbar l-i ha-af 'broke to-me the-nose', koevet l-i habeten 'hurts to-me the-stomach'
Transitive topical object construction	akca oti dvora 'stung me a-bee', acar oti shoter, 'arrested me a-policeman'

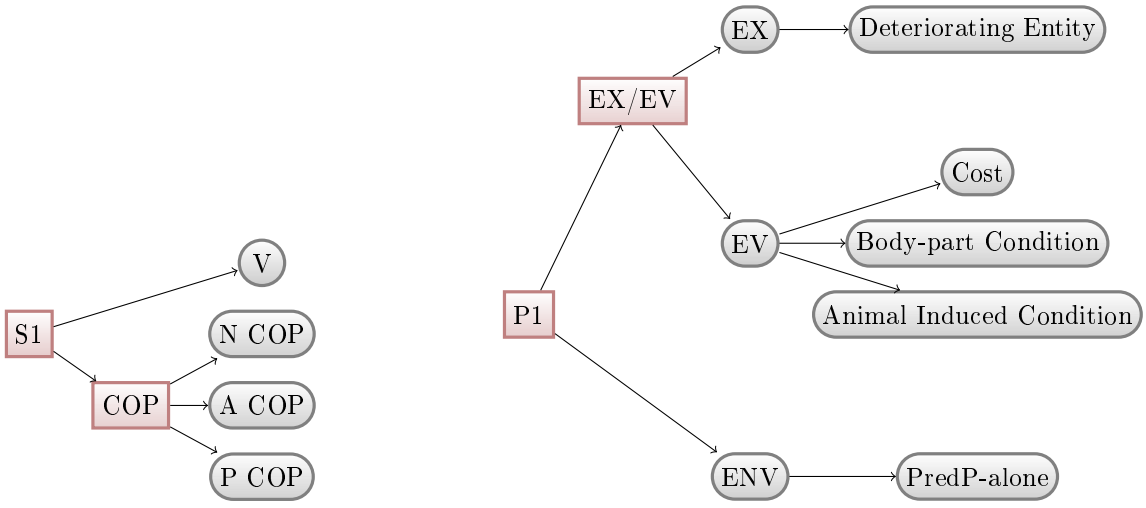

Figure 3: The field of S-patterns in Hebrew.

Table 2 and the discussion below). In the core of this category is the existence predicate *yeš* 'there-is' and it has three peripheral rings.

1. First ring: verbs with a strong existential meaning: *kara* 'happen', *himšix* 'continue', *niš'ar* 'remain', *hofi'a* 'appear', *ba* 'come', *nocar* 'emerge', *ala* 'arise', *hitpate'ax* 'develop', *camax* 'grow'.

2. Second ring:

   (a) verbs of motion that along with a complement and with the construction force of the EX S-pattern acquire the existential-presentational meaning: *azav* 'leave', *nafal* 'fall', *hegi'ax* 'surface'.

   (b) verbs that express intrinsic behavior or state of being of an entity. (*doleket kan nura* 'is-lit here a-light-bulb', *xonot kan kirkarot ec yefeyfiyot* 'are-parked here carriages (of)tree beautiful', *camax bagina perax* 'grew in-the-garden a-flower', *niftax ha-petax* 'was-opened the-opening', *nisdak hasedek* 'was-cracked the-crack').

3. Third ring: Predicates with no existential meaning: the verbs *asa* 'has done' and *ana* 'answered'. See Kuzar's examples in (23):

(23) a. lefi meitav zixroni asa et ha-seret stiven spilberg.
according-to the-best-of my-memory made ACC the-movie steven spielberg.
'If my memory serves me, the movie was produced by Steven Spielberg'.

b. be-exad ha-xiyugim halalu ana li kol seksi.
in-one-of the-dialings those answered to-me a-voice sexy.
'One of my calls was answered by a sexy voice.'

Coming back to the example in (22), we can note that because of the directional complement *le-kivuni* 'in my direction' the event also has an aspect of meaning that relates to appearance and thus to existence. The event in (22), while not existential *per se* can still fall under the second ring of the existential CC, and can thus be expressed by the EX S-pattern.

Kuzar (2006b) further argues that the EX S-pattern is a productive pattern in Hebrew. It seems that whenever a non typical V1 verb is placed in a V1 construction it obtains an existential flavor. Table 2 lists the frequencies of the different V1 verbs in my sample of Linzen's blogs corpus. Except for four verbs, all the frequent verbs were existence verbs (and bear in mind that the most frequent V1 verb, the existence predicate *yeš*, was excluded from my sample because of its fixed word order). Bearing in mind these data and considering that grammaticization is sensitive to frequencies, Kuzar's claim is strongly supported.

Table 2: [V S] verb types and their frequencies

Predicate	Frequency	Function
higi'a 'arrived'	55	existence, presentation
avar 'past'	46	existence
nigmar 'was-over'	28	existence
hitxil 'began'	28	existence
ba 'came'	23	presentation
yaca 'left'	14	existence
nish'ar 'remained'	12	existence
halax 'went'	9	existence
kara 'happened'	8	existence
nixnas 'entered'	8	presentation
histayem 'was-over'	7	existence
nocar 'was-created'	6	existence
yashav 'sat'	6	existence
yarad 'went-down'	6	existence
ala 'rose, went-up'	5	existence, presentation
nishbar 'broke'	3	change of state
nafal 'fell'	3	change of state
halax le'olamo 'died'	3	existence
met 'died'	2	existence

The attentive reader might notice that I have labeled verbs such as *halax* 'went' and *yarad* 'went-down' as existence verbs. This is in fact another indication of the validity of Kuzar's arguments. My sample of the Linzen corpus revealed time and time again that verbs that are normally agentive in the S1 word order, can also appear in V1 sentences, but with a non-agentive meaning.

(24) a. halax l-i ha-kol.
went to-me the-voice.
'I lost my voice.'

b. racu l-i hayom tekstim b-a-roš.
ran to-me today texts in-the-head.
'(Different) texts were going through my mind today.'

c. yarad šeleg.
went-down snow.
'It snowed.'

One may argue that the meaning of the above V1 predicates is so distant from the meaning of the corresponding agentive S1 predicates that they should in fact be considered different predicates. While I do not fully share this sentiment, I wish to point out that there are examples where the V1 predicate clearly maintains its original meaning, but also gains an existential flavor from the construction. Consider for example the following context: *ani hekamti et ha-(ictadyon|opera|bama|iton) bi-šnot ha-šmonim* 'I founded the (stadium|opera|stage|news paper) in the eighties.' All the following continuations—all sentences containing canonical unergative predicates in their original meanings but with a strong existential aspect—are clearly felicitous:

(25) a. racu b-o meitav ha-koxavim šel ha-tkufa.
ran in-it finest the-stars of that-period.

'The best athletes of that period ran there.'

b. šaru b-a meitav zamarei ha-tkoufa.
sang in-it finest singers the-period.
'The best singers of that period sang there.'

c. na'amu b-a gdoley ha-am.
spoke in-it finest the-nation.
'The nation's finest spoke there.'

d. katvu b-o meitav ha-katavim šel ha-tkufa.
wrote in-it finest the-reporters of the-period.
'The finest reporters of that period wrote there.'

From the four sentences in (25) it was sentence (25-d) that was attested in Linzen's blogs corpus[16]. The other sentences are similar and are clearly felicitous. These sentences illustrate that when the context strongly sets up a non-subject element as topical (in the case of the above sentences the element *b-o* 'in-it' is of course the topic) even core unergative predicates can appear in the V1 order. The fact that these examples are infrequent, as I explained above, is also predicted by Kuzar, since these verbs are at the third ring of the EX S-pattern and speakers in similar situations will often choose more common existential predicates in their place.

From these examples I conclude, following Kuzar, that given a supportive context, basically any verb can appear in V1 constructions. Conceptual categories can explain both the fact that we can find a-typical action verbs in P1 sentences, and their relative infrequency being peripheral to the relevant CC.

2.3.4 Criticism

The bulk of Kuzar's work is dedicated to describing different sentence patterns, discussing their syntax, semantics and pragmatics. This data driven investigation allows for a much deeper understanding of these constructions. For instance, it predicts the kinds of predicates we typically observe in P1 sentences in a way that explains both their diversity and their relative frequencies. In that respect Kuzar's account is compelling and accurate. I do however differ somewhat from Kuzar on issues of motivation and in particular on the extent in which information structure considerations affect the choice of word order.

Kuzar (forthcoming, p. 168-169) argues that information structure can be said to motivate the choice of word order only with respect to the major sentence patterns: Actions and background states[17] are hinged on a topic and thus require a topic-comment S1 construction, while existence and evaluation do not hinge on a topic and would thus be encoded in a topicless V1 construction. Kuzar however goes on to argue that in the case of minor and non-prototypical situation types this is not the case and that only the situation type itself can directly motivate word order.

Kuzar bases this claim on two types of arguments. Firstly, minor situation types can both advance the plot or deviate from it, they can be construed as either topic-comment or topicless sentences—they are thus not naturally suited to any particular word order. For instance, a possessive statement can be either a link in the topical aboutness chain (e.g. *ani bedicaon. maxar yeš li mivxan* 'I'm depressed.

[16] The actual context for the sentence was *ha-olam ha-ze ke-iton haya meratek* 'ha-olam ha-ze as-a-newspaper was fascinating'.

[17] In this context, Kuzar takes background states to be situation types expressed in copula sentences, i.e. *dan hu yeled tov* 'Dan is a-boy good'.

tomorrow there-is to-me a-test'), or deviate from it by providing background information (*ani lo yaxol lacet axšav. yeš le-mishehu me-ha-avoda yom-huledet* 'I not able to-leave now. there-is to-someone at-work a-birthday'). Secondly, Kuzar notes that SVO languages use different word orders to express the same pragmatic structure. Take for instance example (26) below (adapted from similar examples in Kuzar (forthcoming, p.168-169)):

(26) Q: ma kara lexa? A: yeš li ke'ev roš
Q: what happened to-you? A: there-is to-me a-headache.
'Q: What happened to you? A:I have a headache'.

The inverted Hebrew sentence *yeš li ke'ev rosh 'there-is to-me a-headache'* is a topic-comment sentence in which the topic is the speaker (realized by the dative element *li*) and the comment is the new information added about the speaker, that she has a headache. The same situation and thus the same topic-comment relations are expressed in English in the V S-pattern—*I have a headache*. Kuzar concludes that it is not information structure considerations, i.e. the deviation from topic-comment propositions, that motivate the Hebrew word order in these cases, but rather the situation type itself (or more precisely, the fact that the situation type of possession is mapped into the EX S-pattern in Hebrew and to the V S-pattern in English).

What I believe is being missed here, is that as suggested by Givón (1976a) and Lambrecht (1994, 2000), it is not the autonomous effect of information structure that is so relevant to the choice of word order, but rather its interaction with the grammatical category of subject. If we take the VS word order to be motivated by the need to code non-topical subjects then all of Kuzar's reservations disappear and we are left with a very strong generalization that is valid for major and minor situation types alike. Indeed, if we reexamine Kuzar's common V1 situation types given in table 1 on page 16, we note that in all cases, irrespective of the questions of whether the sentence has a topic or not and whether it advances the story line or deviates from it, all V1 sentences have non-topical subjects. Furthermore, in the case of the sentences in (26) I will argue that it is the choice of subject that is different between Hebrew and English[18]. Once the subject is selected, the fact that the English sentence model will be S1 and that the Hebrew sentence model will be V1 is fully predictable from our generalization.

Despite these facts, I do agree with Kuzar that information structure can not by itself account for the whole range of data (see chapter 3 and in particular section 3.4 below). I would like to suggest however, that the influence of information structure on the choice of word order is stronger than suggested by Kuzar and is not limited to the prototypical instances of the major situation types. Furthermore, while I agree with Kuzar that ultimately word order is determined by the language specific mapping of situation types into grammatical forms, I suggest that this mapping is probabilistic in nature and that it is best modeled by an approach that takes into account the relative influence of various aspects of the situation (or more precisely, of the proposition describing the situation). Indeed, Kuzar himself notes that the mapping is not fully predictable, among other reasons because of the fact that SV is the unmarked word order and it can accommodate many of the P1 situation types. However, he stops short of providing a comprehensive account of the exact factors that bear on this mapping and of their relative strengths. In that respect the account I'll present in chapters 4 and 5 can be seen as an explication of this aspect of his framework.

[18] In section 3.2 and appendix B, I argue that the grammatical subject is the mechanism language use to uniformly code aspects of propositions that usually manifest themselves in the same sentence element. In that respect English seems to consider animacy to be a key factor in the coding of subjects and Hebrew seems to prefer the cause or source of the eventuality.

3 Inversion as a Low Topicality marker

3.1 Overview

The unmarked pragmatic structure of propositions is topic-comment. Since Hebrew is an SVO language and since the subject for the most part coincides with the topic[19], we get that the unmarked Hebrew sentence is an SVO sentence in which the subject is also the topic. Givón (1976a) argued that when the speaker wishes to convey a proposition in which the subject is not the topic, she will signify it by using a different grammatical form[20]. Hebrew V1 constructions can then be seen as this marked form, and their function is therefore to signify deviance from the unmarked pragmatic structure—to mark non topical subjects[21].

Lambrecht (1994, 2000) used a similar notion in order to motivate grammatical forms cross linguistically. Lambrecht termed the SVO topic-comment sentences (where the subject is also the topic) *predicate focus* sentences[22] and argued that cross-linguistically the need to signify deviation from the unmarked predicate focus structure motivates the availability of different grammatical forms—marked by either intonation or word order—and their association with various linguistic features. Lambrecht suggestion is then similar to Givón's since sentences (with a subject) that deviate from the predicate focus structure will always have non-topical subjects[23]. Lambrecht (1994) discussed two types of pragmatic structures that should be differentiated from predicate focus: (i) sentence focus: a pragmatic structure in which both the subject and the predicate are in focus (and thus not topical. e.g. Q: What's the matter? A: *My* CAR *broke down*), and (ii) narrow focus: a pragmatic structure in which the subject is in focus but the predicate is part of the sentence presupposition (e.g. Q: *what broke down?* A: *my* CAR *broke down*). Lambrecht considered narrow focus and sentence focus to be separate formal categories that can be expressed grammatically in different ways (in English both structures are expressed by the same intonation pattern—a pattern that is different from that of predicate focus sentences). Narrow focus sentences of the type mentioned in the literature are very rare in discourse (this is due to the fact that when faced with a question like *what broke down?* the speaker will normally just reply *my car* and will not repeat the predicate) and so I will not be able to say much on this issue. This discussion does indicate however, that while the Hebrew V1 word order is one construction

[19] The strong statistical correlation between subjects and topics is cross-linguistic (at least in languages that clearly mark subjects). This correlation is not surprising since the function of the subject is, to an extent, to code the topic (see section 3.2 and appendix B).

[20] To be precise Givón favored a scalar concept of topicality over the discrete concept of topic. Givón used the term topicality to refer to a degree of topichood. He also devised a method to measure the topicality of an NP from its textual surroundings but as I will argue in note 68 on page 59 I am not fully confident that his measurement system is in accord with the definitions of topic I will be discussing in this chapter. While I am not in principle against a scalar view of topicality, I will not adopt this approach in this work. The terms topichood and topicality will thus refer to the same thing—the quality of being a topic. In the occasions where I refer to an element as having high or low topicality, the statement should be interpreted as a reference to the probability of the element to be considered the sentence topic (based on its linguistic features, see section 3.5).

[21] Marked word order is one of the two main grammatical forms languages use to code non topical subjects, the other being intonation. Many languages, Hebrew included, use both mechanisms to various degrees. While I do not have quantitative data to bear on this, it can be observed that SVO sentences with non-topical subjects will often involve a deviant intonation pattern where the subject is stressed. Givón (1976a) argued that languages can be put on a continuum with regard to their degree of reliance on both mechanisms: on the one hand English relies mostly on intonation, on the other hand Spanish relies mostly on word order, and Hebrew is in between, combining both mechanisms. Givón further argued that Hebrew is gradually shifting toward a more prevalent use of intonation, but discussion of these facts falls well outside the bounds of this thesis. For a recent comprehensive typological study of these facts see (Sasse, 2006).

[22] The name predicate focus stems from the fact that the predicate, i.e. the verb and its object, are not topical and are in the focus domain.

[23] the only other way to conceive a deviation from predicate focus that does not involve non-topical subjects is if the sentence has a topical subject but no comment. This all-topic sentence model (with no assertion) is not attested in human (adult) language.

that is motivated by the need to mark non-topical subjects, a more fine grained examination of the spectrum of sentences with non-topical subjects might point to other such constructions. Narrow focus constructions are a case in point, but due to the scarcity of data I leave the question of their coding as a topic for further research[24].

Following the above discussion I argue, following Givón, that the pragmatic function directing our use of the Hebrew VS order is the marking of non-topical subjects. This idea can be exemplified with the following passage from a fictional conversation between Dana and her friend:

(27) Dana: kaniti mixnasayim xadašim b-a-kenyon etmol, vekše-xazarti habayta
Dana: I-bought pants new in-the-mall yesterday, and-when-I-returned home
samti otam bi-mxonat ha-kvisa. kše-hitorarti ba-boker badakti
I-put them in-the-machine laundry. When-I-woke-up in-the-morning I-checked
ma šlomam, ve-at lo ta'amini — ha-mixnasayim hitkavcu.
how are they doing, and-you won't believe — the-pants shrank.
'I bought a new pair of pants in the mall yesterday, and when I came back home I put them in the washing machine. When I woke up this morning, I checked on them, and you won't believe it — the pants shrank!'

In the example in (27) the pants are clearly topical by the time we process the final clause. Because of that, using an inverted word order at that point (e.g. *hitkavcu li ha-mixnasayim* 'my pants shrank') would sound awkward compared to Dana's original statement. Furthermore, in a discourse situation where *the pants* are not topical, the inverted word order will sound perfectly natural. See for instance a possible continuation to the passage in (27), this time involving (a gloomy) Dana and her mother:

(28) a. ima: lama at acuva?
Mother: Why you sad?
'Why are you sad?'

dana: hitcavcu li ha-mixnasayim ha-xadavsot.
Dana: shrank to-me the-pants the-new.
'My new pants shrank.'

b. ima: lo šamati tov, ma kara l-a-mixnasayim?
Mother NEG I-hear well, what happened to-the-pants?
'Mother: I didn't hear you well, what happened to the pants?'

dana: hem hitkavcu.
Dana: they shrank.
'They shrank.'

In Dana's reply in (28-a) the pants are not topical; the statement is not perceived as being about the pants but rather as a statement about the speaker, Dana, and the comment is that her pants shrank. This strengthens Givón's claim that inverted word orders mark non-topical subjects. Later still, in

[24] There has been some deliberation in the literature on this issue. Melnik (2002, p. 141-142) argued that in Hebrew narrow focus is not expressed by V1, but rather by intonation (giving as an example her judgment on the sentence Q: *what broke?* A: *HA-AGARTAL nišbar* 'THE-VASE broke'. She also asserted that the reply *nišbar ha-agartal* 'broke the-vase' would be unacceptable in the given context.) Givón (1976a, p. 159) on the other hand argued that narrow focus sentences can be expressed in the V1 order giving as an example his judgment on the sentence Q: *Who gave you the book?* A: *natna li oto ha-xavera šeli.* 'gave to-me it the-girlfriend mine'. It seems to me that we should avoid using introspective sentence judgments when discussing this issue. For the time being I am contend with pointing out the disagreement and deferring conclusions until further research is carried out.

(28-b), when Dana's mother asks her again *what happened to the pants?*, the pants become topical and Dana can only use the SV word order when replying *they shrank* (notice also that the use of a pronoun in Dana's answer makes the VS order completely unacceptable).

In the reminder of this chapter, I will explicate the concepts of subject and topic while providing further evidence for the view that inverted sentences mark non-topical subjects. I will argue however, that while marking non-topical subjects is the central driving force behind the choice of the VS word order, it cannot by itself account for the whole range of data. Only an analysis that considers the simultaneous influence of multiple factors can best account for the phenomenon at hand.

3.2 Subject and Topic

In this thesis, I assume the existence (at least in Hebrew) of the grammatical category subject and the pragmatic category topic. In Hebrew the subject is the element of the sentence that is characterized by agreement with verb and by the nominative case (word order is not a very good indicator of subjecthood in Hebrew, since as this thesis demonstrates, Hebrew subjects can also appear after the verb). This definition equates the subject with the grammatical subject, and I'll be using these terms interchangeably. Following Evans and Levinson (in press), I take the function of subjects to be the uniform coding of various aspects of propositions that typically manifest themselves in the same sentence element (e.g. topicality, agentivity, causality etc.) As for topics, I follow the traditional definition, equating the topic with "what the sentence is about". This concept is admittedly vague but in my opinion it can be partially clarified by using Gundel's definition (Gundel, 1988):

(29) Topic Definition: An entity, E, is the topic of sentence, S, iff in using S the speaker intends to increase the addressee's knowledge about, request information about, or otherwise get the addressee to act with respect to E.

Another way to understand aboutness is through Reinhart's catalog metaphor. Reinhart (1981) compares the speaker and the hearer's representation of the discourse context to a list of propositions they consider true—their context set. Reinhart suggests that in much the same way that library books are indexed by author or title, the propositions in our discourse context are indexed by topic. Once the hearer encounters a new sentence, he identifies its topic and "catalogues" the proposition under its entry in the context set. If the proposition is topicless, it remains uncatalogued (supposedly in a list of topicless propositions). Within this metaphor, the topic is seen as an instruction from the speaker to the hearer to catalogue a proposition under a specific context set entry.

The exact characterization of the subject and topic categories is quite controversial and to a lesser extent so is their use as primitives in linguistic argumentation. An exhaustive discussion of these two concepts is outside the scope of this thesis but the reader is referred to Appendixes B and C and to the references therein for a more in depth discussion of these concepts.

3.3 Inversion as a mechanism to mark non topical subjects

In section 3.1 I argued, following Givón (1976a) and Lambrecht (1994, 2000), that the motivation for the availability of V1 constructions and their association with various linguistic features is the need to

signify deviation from the unmarked topic-comment pragmatic structure where the subject is also the topic—i.e. to mark non topical subjects.

In order to exemplify the potency of this generalization, let us review a representative sample of V1 sentences with different kinds of predicates, subjects and modifiers. If not otherwise specified all examples are from Linzen's blogs corpus.

(30) Sentences involving Existence

a. yeš makot.[25]
EXISTS a-fight.
'There's a fight'

b. yored gešem.
falls rain.
'It's raining.'

(31) Sentences involving Appearance

a. ba ha-menahel ve-amar l-i še-ani ovedet b-a-kupa.
came the-manager and-told to-me that-I am-working at-the-register.
'The manager approached me and said that I'm working at the register.'

b. higi'a ha-pica.
arrived the-pizza.
'The pizza arrived.'

c. ha-rakevet acra. ala aleyha gever ben šišim [...] ve-hityašev mi-cid-i
the-train stopped. climbed on-top-of-it a-man aged sixty [...] and-sat at-my-side
ha-šeni.
other.
'The train stopped. A sixty years old man entered and sat in front of me.'

(32) Sentences involving Change of State

a. nirdam l-i ha-gav.
fell-asleep to-me the-back.
'My back went numb.'

b. kmo be-xol hofa'a tova, [...], nikra l-o ha-meytar.
as at-any concert good, [...], was-torn to-him the-(guitar)-string.
'As at every good concert, he tore his guitar string.'

(33) Sentences that involve a topical object

a. hitxil iti mišehu b-a-mesiba.
flirted with-me someone at-the-party.
'Someone made a pass at me at the party.'

b. helxica oti ha-noxexut šelo.
pressured me the-presence of-him.
'I was pressured by his presence.'

The existential sentences in (30) are all topicless thetic sentences, and are therefore prototypical V1 examples. The reason is that we do not consider an ontological claim of an entity's existence as information about the entity. The speaker in these sentences normally attempts to report an event

[25] http://www.tapuz.co.il/Forums2008/ViewMsg.aspx?ForumId=126&MessageId=1020596&r=1

rather than to provide information about an entity. This argument similarly holds for the appearance sentences in (31). It should be noted however, that appearance sentences will often involve non-topical entities that may or may not become topical later on. We will not normally consider these entities topical in the clause where they were merely presented—mere appearance on the scene does not constitute information about the entity[26]—if in subsequent discourse the speaker provides information about these entities' actions or traits then they'll become topical. As for a subject's change of state situations, when involving a dative element, it is frequently the dative element is topical and not the subject. In these cases the sentences will tend to appear in the V1 order like the sentences in (32). The sentences in (33) exemplify this further—when the object is the topic and not the subject, these sentences will tend to appear in V1 order.

The potency of this generalization is also apparent when we examine the list of common V1 situation types devised by Kuzar (see table 1 on page 16). It is clear that irrespective of the presence or absence of a topic element, in all cases where a V1 sentence has a topic, that topic is not the subject.

This generalization is not entirely without exceptions but it is very robust and as demonstrated by Lambrecht (1994, 2000) it cross-linguistically plays a key role in motivating grammatical forms. In the next section the limitations of this generalization will be discussed and I will argue that despite its key role in motivating V1 constructions, it cannot account by itself for the "online" choice of word order.

3.4 Why topicality is not enough

In section 3.3 I have examined common V1 sentences and it was evident that in all sentences the subject was not the topic. However, we may not conclude that topicality is all that is required in order to account for the choice of word order. The SVO order, being the unmarked word order, can often accommodate non-topical subjects, and furthermore, to a lesser degree, topical subjects can appear in the V1 word order. Consider the sentences below from Linzen's blogs corpus (the subjects are in bold):

(34) a. **merc** mistayem.
march is-ending.
'March is coming to an end.'

b. **maxšavot** racu.
thoughts ran.
'Thoughts were running (through my mind).'

c. ha-ben šel-i omer l-i: "aba, **mišehu** herbic l-i."[27]
the-son mine is-saying to-me: "dad, someone hit to-me."
'My son is telling me: "dad, someone hit me."'

The sentences in (34) are all S1 sentences where the subject is not the topic. Sentence (34-a) is a thetic topicless sentence that reports a background situation; sentence (34-b) is again thetic, this time the topic (the speaker) is not mentioned in the sentence; and in sentence (34-c) the topic is the speaker (the dative element *l-i* 'to-me') and so again, the subject is not topical (as is also verified by examining

[26] One can also think of it in terms of the catalog metaphor. For instance in (31-a) the clause *ba ha-menahel* 'came the-principal' just signifies a possible future topic and perhaps opens a catalog entry for it, we would not tag the fact that the principal arrived under his entry. Later, if information is given about his actions it will be labeled under his newly created catalog entry and at that point he will become topical.

[27] "Born again" forum: http://sc.tapuz.co.il/shirshurCommuna-8765-3365926.htm

the continuation of that discourse). The choice of word order in all these sentences does not stem from the subject being topical, but rather from different factors. First of all, the S1 word order is unmarked so it can accommodate a wider range of situation types than the V1 word order. Beyond that however, other factors are at work. In sentence (34-a) it may be the influence of the present tense[28]. In sentence (34-b) there are really not that many factors that support the choice of the S1 word order. From the set of factors I will consider in part II, only the subject's length (1 word) favors the choice of the S1 order. I suspect that in this case the choice was either due to pure chance or to the idiosyncratic properties of the verb[29]. In sentence (34-c) the choice of word order can be attributed to numerous factors: the subject's animacy, the verb class (an unergative, agentive verb) and the NP length (1 word) are all factors that favor the S1 word order and that can account for the word order choice in this case.

The following example shows that even the generalization that V1 sentences code low topicality is not without exceptions (the subject of the relevant sentence is in bold).

(35) nixnas porec ha-bayta, maca et ha-maftexot al ha-šulxan ve-lakax et ha-oto. halax
entered a-burglar home, found acc the-keys on the-table and took acc the-car. went
ha-oto.[30]
the-car.
'A burglar entered my house, found the car keys on the table and took the car. The car is gone.'

In sentence (35) (i.e. *halax ha-oto* 'went the-car') the subject *ha-oto* 'the car' is clearly topical. The text is from a report about a conversation between a client and his insurance company. The car is the discourse topic, it was mentioned in the immediately preceding sentence, it is definite and strongly topical. The choice of the V1 word order can stem in this case from a combination of the non-animacy of the subject, the verb class (the verb in this context expresses (non) existence/change of state) and probably from the idiosyncratic properties of the verb *halax* (in its existential meaning) that appears to favor the V1 word order to an even larger degree than other unaccusative verbs.

3.5 Topic Hierarchies

Over the years, numerous studies converged on a large group of grammatical features that appear to correlate with topicality. These studies investigated various phenomena ranging over a highly diverse language base involving Semitic, Bantu, Slavic, Germanic and Romance languages (cf. Hawkinson and Hyman, 1974, Timberlake, 1975, Givón, 1976c,a, 1983, Comrie, 1981, Lambrecht, 1994, inter alia).

[28] I did not statistically model the effect of tense on word order as it only became apparent to me in later stages of my work. However, I did find a reference to this influence in the work of Shlonsky (1987, p. 143) who argued that the present tense lends verbs a more habitual and continuous aspect that makes them less appropriate for V1 sentences. From the perspective of topicality it does seem reasonable that habitual events will be associated with high topicality (discussing the habitual action/behavior of a non-topical entity seems somewhat unlikely to me), but I did not encounter any research on this issue. Naturally in the context of (34-a) the reported event is not habitual, but still, it is possible that due to its effect on meaning the present tense as a whole became somewhat disfavored in V1 sentences. The exact influence of tense on word order should be further researched.

[29] in this context the verb expresses existence, which is a property of V1 sentences. However, some verbs behave differently from others even within the same semantic group (i.e. some existence verbs prefer the V1 order more than others, etc.) This aspect of idiosyncratic meaning can also be modeled statistically by taking into account the specific verbs involved, but it requires a larger corpus that the one I used in this study.

[30] blog post: http://www.yr.co.il/blog/index.php?m=200903&paged=2

Every such feature represents a hierarchy—the higher the entity is in this hierarchy, the higher is its probability to be considered the topic. Table 3 lists the various features along with their values[31].

Table 3: Topic Hierarchies

Feature/Hierarchy	Feature Values
Person	1st > 2nd > 3rd
Animacy	Human > Non-Human
Definiteness	Definite > Indefinite
Thematic Role	Agent > Benefactor > Patient
Accessibility	Old > Given > New
Subject Coding	Pronoun > Lexical NP
NP Size	Light > Heavy
Verb Class	Unergative > Unaccusative

The different hierarchies are a result of typological research, but the association of all factors with topicality is also very intuitive: people tend to talk about themselves more than they talk about other people—hence the person hierarchy; they tend to talk about people more than they talk about inanimate objects—hence the animacy hierarchy; they tend to talk more about people who are performing actions rather than entities receiving actions—hence the thematic hierarchy; they talk more about entities their listener has in mind then of new entities (and when they want to talk about new entities they usually first introduce them and only then discuss them)—hence the accessibility hierarchy. The subject coding hierarchy can be derived from accessibility (cf. Ariel, 1988, 1990, 2001) and so can NP size.

It is interesting to note, that while the precise definition of topic remains controversial, the relevance of topic hierarchies to various grammatical phenomena is well established and robust. Because of the difficulty to define topic in a precise non-intuitive way, we can use the correlates to show its influence. If our assumption that the function of V1 sentences is to mark non-topical subjects, then we would expect that V1 constructions will lean toward the non-topical edge of all the topic hierarchies listed in table 3. Furthermore, since grammaticization is sensitive to frequencies, we would also expect that after a while the above non-topical features will become (at least softly) grammaticized and associated with V1 constructions. At that point the choice of word order will not be influenced only by the conceptual function of the construction, but rather by all the above factors, to various degrees. In chapter 4 and 5 I will present results that show that this is indeed the case—V1 constructions lean more toward the non topical edge of all the topic hierarchies than S1 constructions, and furthermore, only the combination of multiple factors predicts the choice of word order in an optimal manner.

[31] Topic hierarchies deal with the topicality of entities and that is why all relevant features but one are features of NPs. The addition of the hierarchy for verb class is my own but it is an immediate by-product of the hierarchy of thematic roles (a subject who is an agent normally entails an unergative verb whereas a subject who is a patient normally entails an unaccusative verb).

3.6 Discussion and Concluding Remarks

In this chapter I have argued based on qualitative data and on the existing literature that the driving force behind the choice of word order is the need to mark non topical subjects. I have argued that (i) the need to mark non topical subjects accounts for the association of different linguistic features with V1 constructions, and (ii) that since grammaticization is sensitive to frequencies these features may now influence word order themselves and only weighing their relative influence will best account for word order data. In part II I will present a quantitative analysis that will confirm that (i) all features that correlate with topicality indeed correlate with word order in the expected manner, and (ii) that in order to best account for word order data multiple weighted factors have to be taken into account.

There is however one remaining gap between the arguments set forth in this paper and the results obtained: even if the empirical results are accepted, there is no compulsion to accept that it is the need to mark non topical subjects that motivates the association of the different factors with V1 constructions. The factors above all pattern the same way, so just as I have taken topicality as the overarching organizing principle and used it to explain other hierarchies, one might select a different factor (e.g. accessibility) as the general principle behind the hierarchies and construct a similar argument.[32] Another account may avoid a unifying concept altogether. As discussed in section 3.2 (see also appendix B.2) I take the subject to be a grammatical mechanism to code propositional aspects that correlate statistically. It can then be argued that the topic is just one factor that along with the other factors in section 3.5 can affect the coding of an element as the subject. When the interactions between the factors are such that the element is less "subject like"—this is marked by word order. This account might actually be less speculative than the account presented here, since it does not make any assumptions that are not backed up by quantitative analysis. Nevertheless, based at this point on my own intuitive observations, as well as on existing work that supports a similar point of view, I would like to suggest that topicality does have a privileged role in the existence and availability of grammatical forms (and word orders) and in the association of the other factors with them. Below are my reasons for the above viewpoint:

1. Marking non topical subjects is sensible from a discourse perspective.
   If the hearer encounters a canonical subject (i.e. characterized by the triplet of case, agreement and word order) he will tend to assume it is topical and process it as such (e.g. associate the proposition with it, see Reinhart's metaphor in section 3.2). If that subject is not topical this may result in misallocated attention and can disrupt discourse. That goes to say—coding non topical subjects is something that needs to be done.

2. Word order alternations may be required more in the marking of non-topical subjects than in the marking of other linguistic features.
   The semantic characteristics of an entity can normally be derived from its lexical entry (e.g. animacy) or from compositional semantics (e.g. agentivity). Likewise, marking accessibility is to a large extent achieved by the entity's NP form (cf. Ariel, 1988, 1990, 2001). For marking non topical subjects, languages have developed different means, and it seems that word order is one of the more common ones (cf. Sasse, 2006, for typological data). That goes to say—the values of the other features I considered here can be inferred quite naturally by other means, so they are less likely to require separate grammatical forms as their coding mechanism.

[32] admittedly, the argumentation becomes less natural with other factors. Try for instance linking accessibility with verb class

3. From my own impressions of the relevant data in the literature, and from initial analysis of my corpus, the subject's topicality (or non topicality) seems to be a stronger factor than the others in the prediction of word order (see also note 68 on page 59 regarding initial corpus evidence). As argued in section 3.4 it can not synchronically account for the whole range of data, but still, it is probably the most influential factor. Furthermore, V1 sentences with clearly topical subjects appear to be harder to come by than V1 sentences with any of the other "non topic like" factor levels. I qualify these statements however, since my own intuitions of topicality and aboutness are likely to be influenced by my data and by my theoretical bias. Obviously a study that shows the ability of non-linguist informants to agree on intuitions of topicality is needed before empirical conclusions can be drawn. Such a study is currently beyond the scope of this thesis and it should be a subject for future research.

4. Taking marking of non topical subjects to be the motivation for word order alternations is consistent with mainstream functional linguistics and backed up by cross linguistic research (cf. Givón 1976b, 1983, Lambrecht 1994, 2000 and the references in section 3.5).

Due to the above reasons it is reasonable to consider the marking of non topical subjects to be motivating the availability of word orders and their association with different factors and factor levels. I concede however that at this point the suggested motivation is just one plausible account that is in accord with our data. In addition to the alternatives discussed above, Gries (2003) has used a different approach in his discussion of English particle placement. In his study, Gries (2003) examined the influence of 21 factors on the speaker's choice of verb–particle construction. The two alternating constructions he considered were *the continuous construction* (e.g. *John picked up the book*), and *the discontinuous construction* (e.g. *John picked the book up*). Among the factors examined in both this thesis and in Gries's study, all patterned the same way: factor levels that were associated with the Hebrew subject–verb word order were associated with the English discontinuous construction whereas factor levels associated with the Hebrew verb–subject word order were associated with the English continuous construction. Similar results were obtained in another recent study of syntactic variation—Bresnan's multifactorial study of the English dative alternation (Bresnan et al., 2007)[33]. With regard to English particle placement Gries (2003) suggested two cognitive motivations for the obtained results: first he suggested (Gries, 2003, chapter 4) that the correlations of factor levels and word orders may result from processing considerations where the speaker attempts to minimize processing effort for both himself and the hearer; then he examined a number of connectionist models and suggested that the correlations may stem from principles of spreading activation in neural networks. Gries's account in terms of spreading activation is especially appealing since it relates directly to low-level cognitive information processing mechanisms. In this respect it is interesting to note Deane's cognitive interpretation of topics (Deane, 1992, p. 36-38, 187-194) according to which topical elements are sentence elements whose salience is due to spreading activation whereas focal elements are elements whose salience is (conversely) due to cognitive focus. Indeed, if the term topic is interpreted this way (rather than in the more traditional linguistic sense I presented here), Gries's cognitive account and the account presented here may very well coincide.

[33] Bresnan et al. (2007) examined the influence of 14 factors on the speaker's choice of dative construction. The two alternating constructions considered were the prepositional dative construction (e.g. *John gave the book to Mary*) and the double object construction (e.g. *John gave Mary the book*). The results obtained were that the prepositional dative construction exhibits a preference for direct objects that are animate, definite, accessible, pronominal and short. The same is true for the preferred subject of the Hebrew subject–verb word order and the preferred direct object of the English (verb–particle) discontinuous construction.

Due to my limited qualifications in the field of cognitive science, I will not attempt to evaluate these approaches here. I will rather maintain the traditional concept of topic, and leave open the question of its cognitive manifestations. I would like however to suggest a future research paradigm that can be employed in order to falsify or improve upon the topicality account. The initial step of such research should be putting forth a factor (other than topicality) that may account for the association of Hebrew V1 constructions with the various features discussed in section 3.5. As I have suggested above, a number of such factors are already known. Beyond that point however, the prospective researcher should also: (i) provide a reason for why it makes sense for the factor to be grammatically marked (as I discussed above, it makes less sense to use a marked grammatical form to mark features that can be to a large extent inferable by other means); (ii) show that the factor is likely to be motivating word order cross linguistically; and optimally (iii) differentiate their account from the one presented here by pointing out different factors that are associated with the two accounts and by providing quantitative evidence that a model that contains one set of factors is stronger than a model that contains the others.

Part II

Empirical Analysis

4 Data Collection and Analysis

4.1 Methodology and Experimental Hypothesis

In this chapter and the next I will present the results of a quantitative corpus investigation designed to bear out the argument outlined in part I, as summarized in section 3.5. Two groups of sentences are examined: one that exclusively includes verb initial sentences (the V1 group) and one that exclusively includes sentences that open with an NP subject that is followed by a verb (the S1 group). My two predictions following the part I discussion are outlined in (36) and will be referred to as "the topicality hypothesis".

(36) The Topicality Hypothesis

a. All factors that correlate with topicality will also correlate with word order and they will do so in the following manner: factor levels that align with high topicality will align with the S1 word order and factor levels that align with low topicality will align with the V1 word order.

b. No single factor can exclusively account for the facts of word order. The data is best accounted for by a set of factors—irreducible to one another—all of which make an independent contribution to the choice of word order.

In accord with (36-b), a further goal of these two chapters is to arrive at a set of significant irreducible factors that influence the choice of word order.

In the continuation of this chapter I will discuss the corpus used in this study and the manner in which it was analyzed. In chapter 5, I will use statistical methods to bear out the topicality hypothesis. I will use monofactorial analysis in order to support prediction (36-a) and multifactorial analysis to support prediction (36-b). All statistical analysis in this work was carried out using the free and open source R project for statistical computing (R Development Core Team, 2009).

4.2 Data Origins

The corpus used for the quantitative part of this work is Linzen's blogs corpus (Linzen, 2009). It includes blog posts by various bloggers writing in different genres and registers. It is the largest corpus for Hebrew texts and contains more than 50,000,000 tokens.

I have used Melingo's part of speech tagger to tag the corpus[34] and then programmatically extracted a group of subject initial sentences and a group of verb initial sentences. The automatic extraction was followed by a lengthy manual process in which I randomly selected sentences from the two groups while making sure that all chosen sentences contain a subject, a finite verb and no direct or clause objects.

[34] The part of speech tagger was provided to me courtesy of Melingo Ltd. The corpus was also independently tagged for parts of speech by its compiler (Tal Linzen) and the tagged version is now available from him upon request.

In the selection process, I excluded sentences that featured the existence predicate *yeš* due to their high frequency and fixed word order. All selected clauses were simple (i.e. not embedded) although I did include clauses that were part of a clause conjunction. Since this study deals with colloquial Hebrew I have excluded sentences taken from blog posts with a more literary style. This process resulted in a group of 370 V1 sentences and a group of 191 S1 sentences for a total of 561 corpus sentences.

4.3 Factors and Factor Levels

In order to provide evidence supporting the topicality hypothesis, I had to take the factors discussed in section 3 and define them formally so as to allow empirical analysis. In addition to these factors I have also considered the factors Case and Agr (discussed below) that are often mentioned in the context of inverted sentences. These factors are problematic in that they are unlikely to be available to the speaker when she is making her decision about the choice of word order (and for this reason they will not be included in the multifactorial analysis), but I still wanted to examine their interactions with word order. In this section I will review the various factors and provide the operational definitions according to which they were analyzed.

I will open with the group of morphosyntactic factors. This group contains five factors: NPType, Def, Person, Agr and Case.

The factor NPType is a nominal factor representing the type of the subject NP. It has three levels: *Pronoun*, *Proper Name* and *Lexical NP*. Originally I also designated a separate level for kin terms[35] but my data was too sparse to warrant this level. I finally decided to somewhat arbitrarily include them within the group of lexical NPs. As a precaution, I also verified that classifying them as proper names does not change the final results in any significant way.

The factor Def is a nominal factor representing the definiteness of the subject NP. It is given the value 1 if the subject is definite and 0 otherwise. I generally take bare NPs that resist the definite article to be intrinsically definite (e.g. pronoun and proper names).

The factor Person is nominal and corresponds to the person marking of the subject NP. It gets the value 1 for first person, 2 for second person and 3 for third person.

The factor Agr is nominal and is given the value 1 if the subject agrees with the verb and 0 otherwise.

The factor Case is nominal and has the two values—*Nominative* and *Accusative*—depending on the subject's case. If the subject is marked with the accusative marker *et* I take its case to be accusative, otherwise I take it to be nominative.

The group of semantic factors includes the three factors: Agentivity, Animacy and VClass. I will discuss them in turn.

The factor Agentivity is nominal and corresponds to the agentivity of the subject. It is given the value 1 if the subject is agentive and 0 otherwise. A subject is considered agentive if the action described by the sentence is perceived as carried out with the volition of that subject.

[35] Kin terms sometimes behave like a proper names and sometimes like lexical NPs. For instance in a sentence like *ima amra l-i lištof yadayim* 'mother told me to-wash my hands' we cannot add the definite article to the kin term *ima* 'mother'—it appears to be intrinsically definite. This behavior is similar to that of a proper name. However, in examples like *ha-ima šel dan amra l-i lištof yadayim* 'the-mother of dan told me to-wash my-hands' or *ha-ima-ha'xaruca azra kol yom le-bnoteyha be-ši'urey-ha-bayit* 'the-diligent-mother helped every day to-her-daughters with-(their)-homework' the behavior of the kin term *ima* 'mother' is more akin to that of a lexical noun.

The factor ANIMACY is nominal and corresponds to the animacy of the subject. It is given the value 1 if the subject is animate (human or animal) and 0 otherwise.

The factor VCLASS is nominal and corresponds to the type of the clause's verb. It receives one of three values—*Passive*, *Unaccusative* or *Unergative*. The classification of intransitive verbs into the unaccusative and unergative verb classes is not always straightforward. While it is generally agreed that such classification is possible based on the verb's semantics, theories differ with respect to the nature of the semantic traits involved and thus on the manner in which this classification is to be carried out. In a previous work (Taub-Tabib, 2007) I compared several theories of unaccusativity with respect to their ability to predict the Hebrew subject–verb word order. Based on that study and on considerations of clarity and suitability for an empirical investigation Levin and Rappaport Hovav's classification guidelines (Levin and Rappaport Hovav, 1995, p. 135-166) were selected as the basis for verb class classification in this work. The guidelines are phrased in the terms of linking rules that based on the verb's meaning determine if its subject is to be linked internally (i.e. the verb is unaccusative), or externally (i.e. the verb is unergative). The rules are evaluated in order, so once one of them applies, the others are not evaluated and do not affect classification[36].

(37) a. The Directed Change Linking Rule: The argument of a verb that corresponds to the entity undergoing the directed change described by that verb is its direct internal argument.
b. The Existence Linking Rule: The argument of a verb whose existence is asserted or denied is its direct internal argument.
c. The Immediate Cause Linking Rule: The argument of a verb that denotes the immediate cause of the eventuality described by that verb is its external argument.
d. The Default Linking Rule: An argument of a verb that does not fall under the scope of any of the other linking rules is its direct internal argument.

While the classification rules cannot be fully understood without careful reading of Levin and Rappaport Hovav's work, it should be noted that their classification scheme is accompanied by a comprehensive list of preclassified verbs that ease the work of classification considerably (Levin and Rappaport Hovav, 1995, Chapter 3 and Appendix A). Indecisiveness can then only arise when considering a verb that is outside of this preclassified group and even then, in most instances, the verb's meaning would be at least partly analogous to one of the preclassified verbs. A thorough understanding of the linking rules along with the list of preclassified verbs allows for relatively unambiguous analysis.

In the domain of discourse pragmatics, I have considered the factor ACCESS which is nominal and corresponds to the accessibility of the subject NP. I take the accessibility of a noun phrase to be, roughly speaking, the estimated degree of cognitive salience of the representation of NP in the hearer's mind. From the a processing perspective it can be looked at as the degree of effort required to access the representation of the referent the NP stands for.

Quantifying the degree of cognitive accessibility by observing raw texts is notoriously difficult and can only be approximated to an extent. Empirical studies have used different techniques for this purpose. Simple measurements can be obtained by counting previous mentions or counting the distance in words from the last mention (cf. Givón, 1983). The problem with these measurements is that they downplay the role of contextual priming. Some entities, while not directly mentioned in previous discourse, are

[36] The actual ordering of the rules is a,b>c>d. Levin and Rappaport Hovav did not determine the order between the directed change linking rule and the existence linking rule. Note however, that since both rules link the argument internally, their ordering is irrelevant with respect to verb classification.

primed by other preoccuring entities and can be quite salient. Ariel (1990) took these difficulties into consideration, but the accessibility scale she devised is comprised of fifteen levels and is too detailed to be used in this study. Arnold et al. (2000) based on Prince (1981, 1992) addressed contextual priming by using just three levels of analysis: *Given*, *Inferable* and *New*: an entity is *Given* if it appeared in previous discourse, *Inferable* if it was triggered by an entity in previous discourse and *New* otherwise. A notable difference between Arnold's coding system and Prince's original proposal is that Prince did not take *Given* and *New* to be primitives but rather drew a distinction between discourse givenness/newness and hearer givenness/newness. This distinction is important since some entities might be new discourse wise but old hearer wise due to the hearer's world knowledge. Such entities are for example well known individuals, countries, cities etc. It is reasonable to assume that these entities are salient in the hearer's long term memory and thus more easily processed than other *New* entities. For this reason I introduced the level *LTM* (stands for Long Term Memory) which indicates the assumed salient presence of the (otherwise new) entity in the hearer's long term memory.

A final note regards my attitude toward inferables. Prince (1992, p. 9) defined inferables as otherwise discourse new entities that uphold two conditions:

1. The hearer has the belief that the entity in question is plausibly **related** to some other 'trigger' entity, where the trigger entity is itself not hearer new.
2. The hearer should be able to infer the existence of the entity in question.

However, such an inclusive definition is too vague to allow an empirical investigation. To make this definition more concrete one must precisely characterize the possible relations between the entity and its trigger. Recent empirical studies (Michaelis and Francis, 2007, Bresnan et al., 2007) used a criteria of superset mention (a flower can be a trigger for a rose), and Gries (2003) used a metric for cohesiveness that considered, in addition, the relation of part–whole. In addition to these relations I have also considered the relations action–result and subset listing[37]. Table 4 summarizes and exemplifies the levels of Access.

[37] By subset listing I mean the situation in which an item is triggered by a previously mentioned item such that both items are part of the same whole, but the whole in itself is not mentioned. For instance the entity *the teacher* can trigger the entity *the students* even if an entity such as *the class* or *the school* is not previously mentioned.

Table 4: Examples of analysis for ACCESS

Level	*Comments*	*Attested Corpus Examples*
Old	entities that were either mentioned in previous discourse or that are known to the hearer because of the speech setting (e.g. the speaker or hearer)	hitnagašti be-cvi. halax ha-cvi. I-ran-into a-deer, was-gone the-deer. le-zxuto-šel-peter ye'amer še-hu in-peter's-favor it-should-be-said that-he lo šamen-becura-maxli'a [...] hu gam NEG obese [...] He is-also mitlabeš lo ra. dressed NEG badly.
Inferable	entities primed by a preceding trigger.	aval nastia lo hictalma, rak ani, ve-yac'u tmunot yafot. 'but Nastia NEG was-photographed, only I, and-turned-out pictures beautiful.' gisi xole, nafla alav shapa'at. 'my-brother-in-law is-sick, fell on-him the-flu.'
LTM	entities that the speaker assumes can be readily retrieved from the hearer's long term memory.	hilary svank zaxta 'Hilary Swank won' (the oscar). avram grant hitpater mi-'imun nivxeret-yisrael. 'Avraham Grant quit from-training the-israeli-team.'
New	entities that do not fall under the three other categories, i.e. entities that are new to the hearer.	ba'u šney banot xadašot le-beyt-sefer. 'came two girls new to-school'

Finally I also examined the factor LENGTH which is measured on a ratio scale and corresponds to the length in words of the subject NP.

The expected behavior of each of the factors and factor levels are in accord with the topic hierarchies of section 3.5. They should be statistically interpreted as follows: if for a factor X and its levels (x_1, x_2) we predict that $x_1 > x_2$ then: (i) the level x_1 should be more prevalent in the S1 sentence group than expected by chance and less prevalent in the V1 group than expected by chance; and (ii) the x_2 level should be more prevalent in the V1 sentence group than expected by chance and less prevalent in the S1 group than expected by chance[38]. Less formally, the level x_1 should be significantly more prevalent than the level x_2 in the S1 group and significantly less prevalent than x_2 in the V1 group.

[38] To be mathematically precise, If the factor X has more than two levels (i.e. levels $(x_1, x_2, ..., x_n)$, $n > 2$, *where* $x_1 > x_2 > ... > x_n$), then we expect that for every two levels $x_i > x_j$, $1 < i, j < n$ and for a contingency table crossing **just these two levels** with SVOrder, than: (i) the level x_i should be more prevalent in the S1 sentence group than expected by chance and less prevalent in the V1 group than expected by chance; and (ii) the x_j level should be more prevalent in the V1 sentence group than expected by chance and less prevalent in the S1 sentence group than expected by chance. This somewhat cumbersome formulation is the formal way of saying that x_i should be significantly more prevalent than x_j in the S1 sentence group and x_j should be significantly more prevalent than x_i in the V1 sentence group.

The different factors, their levels and their expected behavior with regard to word order are summarized in table 5.

5 Results

5.1 Overview

In this chapter I will use monofactorial and multifactorial analysis techniques in order to bear out the topicality hypothesis as formulated in section 4.1. In section 5.2 I will discuss the first part of the hypothesis—(36-a). I will review each factor in turn, examine its distribution relative to SVOrder and determine whether it meets the expectations as formulated in table 5. I will also equate each factor level with a correlation coefficient indicating the strength of its relation to word order. In Section 5.3 I will discuss the second part of the hypothesis (36-b) and examine the simultaneous effects of all factors in an attempt to determine which factors affect word order in a manner that is not reducible to the effect of other factors. I will examine several multifactorial models and determine a model and set of factors that account for the observed word order data in an optimal manner.

Before I present the results, a caveat is in order. As explained in section 4.2, the data used in this study were randomly selected from two different groups of sentences, the S1 group and the V1 group. The proportions of these two groups—370 V1 sentences vs. 191 S1 sentences (roughly 2:1)—do not reflect their proportions in real discourse. As a result, the models presented in this chapter fall short of providing predictions about the actual probability of choosing one word order over another in natural discourse. The probabilities presented should all be interpreted as the likelihood of choosing one word order over another from a corpus with a roughly two to one proportion of V1 vs. S1 sentences. It should be stressed however, that in light of the goals of this study this issue not problematic. Prediction of word order in natural discourse may be an interesting problem from the standpoint of NLP[39], but it is not the main concern of the linguist. Linguistically, we are interested in understanding the factors that affect word order, their relative strength and interactions. All these data remain the same irrespective of the V1 vs. S1 proportions of the corpus.

5.2 Monofactorial Results

5.2.1 Morphosyntactic Factors

NP Type Since NPType is the first factor to be addressed, I will discuss the statistical procedures involved in its analysis in some depth. The analysis process for the other factors is similar, so subsequent discussion will not reiterate this process and will be limited to listing the results.

The data regarding the distribution of NPType relative to SVOrder is summarized in Table 6.

[39] NLP is an acronym for Natural Language Processing. This is a subfield of artificial intelligence that is concerned with real world applications that relate to language processing. Such applications include (but are not limited to): machine translation, text to speech, speech recognition, speech generation, named entity recognition, information retrieval, etc.

Table 5: Specification of the considered factors and their levels

Factor	Levels	Comments and expectations
ACCESS	**Old** **Inferable** **LTM** **New** See table 4 for definitions and examples.	4 levels, nominal scale *Old* > *Inferable*, *LTM* > *New*
NPTYPE	**Pronoun** **Proper Name** names of people or places **Lexical NP**	3 levels, nominal scale *Pronoun* > *Proper Name* > *Lexical NP*
AGENTIVITY	**0** The subject is not agentive **1** The subject is agentive	2 levels, nominal scale *1* > *0*
DEF	**0** The subject is not definite **1** The subject is definite	2 levels, nominal scale *1* > *0*
PERSON	**1** first person **2** second person **3** third person	3 levels, nominal scale *1* > *2* > *3*
ANIMACY	**0** the subject is inanimate **1** the subject is animate (human or animal)	2 levels ,nominal scale *1* > *0*
CASE	**Nominative** **Accusative**	2 levels, nominal scale *Nominative* > *Accusative*
AGR	**0** The subject does not agree with the verb (in either person gender or number) **1** The subject agrees with the verb	2 levels, nominal scale *1* > *0*
VCLASS	**Unaccusative** **Unergative** **Passive**	3 levels, nominal scale *Unergative* > *Passive*, *Unaccusative*
LENGTH	The number of words of the subject NP.	ratio scale Len(S1 Subjects) < Len(V1 Subjects)
SVORDER	**SV** Subject preceding the verb **VS** Verb preceding the subject	The predicted variable. 2 levels nominal scale

Table 6: Distribution of NPTYPE relative to SVORDER

	V1 Constructions	*S1 Constructions*	*Totals*
Lexical NP	356 (≈83%)	75 (≈17%)	431
Proper Name	14 (≈29%)	25 (≈71%)	39
Pronoun	0 (0%)	91 (100%)	91

Each row in the table corresponds to a level of NPTYPE and it shows the number of sentences of the specified level that appeared in the V1 vs. the S1 sentence groups. The percentages that appear in brackets after every count represent the proportion of V1 vs. S1 sentences of the specified level. It is crucial to keep in mind that the corpus contains ≈66% V1 sentences (370/561) and ≈34% S1 sentences (191/561). Under the null hypothesis of no relation between NPTYPE and SVORDER the expectation is therefore that the total number of sentences of each factor level will pattern the same way—≈66% of them should be V1 sentences and ≈34% of them should be S1 sentences. For example, if we examine the level *Lexical NP,* we note that 431 of the corpus sentences belong to this level (i.e. 431 sentences out of the total 561 had lexical subjects[40]). We therefore expect that under the null hypothesis of no relation between NPTYPE and SVORDER ≈66% of them—approximately 284 sentences—will have the V1 word order, and ≈34% of them—approximately 147 sentences—will have the S1 word order. However, the observed results of ≈83% V1 sentences and ≈17% S1 sentences indicate that sentences with lexical subjects appear more often than expected by chance in the V1 group and less often than expected by chance in the S1 group. As I will soon show, this difference is statistically significant.

Knowing the total number of sentences for each level of NPTYPE, we can repeat the above process for all the levels and arrive at the matrix of expected frequencies outlined below in table 7.

Table 7: Expected distribution of NPTYPE relative to SVORDER

	V1 Constructions	*S1 Constructions*	*Totals*
Lexical NP	284.26 (≈66%)	146.74 (≈34%)	431
Proper Name	25.72 (≈66%)	13.28 (≈34%)	39
Pronoun	60.02 (≈66%)	30.98 (≈34%)	91

The first statistical test we'll use calculates the χ^2 statistic for the overall distribution from tables 6 and 7. Under the null hypothesis of no relation between NPTYPE and SVORDER, this statistic is distributed χ^2 with 2 degrees of freedom. Calculating this statistic yields the value $\chi^2(2) = 245.15$ which is highly significant ($p<0.001$). This means that the null hypothesis is false and that there is a

[40] It should be kept in mind, that because of the caveat discussed in section 5.1, one cannot conclude that sentences in natural discourse contain roughly ≈77%=431/561 lexical subjects. The corpus used in this experiment is skewed towards V1 sentences. Had the corpus contained a more balanced proportion of V1 vs. S1 sentences than the percentage of the sentences with lexical subject would decrease.

correlation between between NPTYPE and SVORDER. Next we'll calculate the cramer V correlation coefficient to determine the strength of this relation[41]. Cramer V for the overall distribution is V=.66 which indicates a very strong correlation.

Now that we know that NPType and SVOrder are strongly correlated, our next step is to determine which of the factor levels (or table cells) contributes significantly to this correlation. As pointed out by Givón (1992) and Gries (2003) it is feasible for certain factor levels to differ significantly from their expected frequencies, while at the same time for other levels of the same factor to fall within the expected range. It is thus misleading to say that the factor is correlated with word order without specifying which of its levels are responsible for this correlation. The procedure I will be using to determine the cells that differ significantly from their expected frequencies is taken from Gries (2003, p. 86). It involves conducting six post-hoc χ^2 tests—one for each cell. Since we know the observed and expected frequencies for each cell, calculating the χ^2 value for the (i,j) cell follows the normal formula for x^2 (i.e. $\chi^2_{ij}(1) = \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$; $1 < i < rows$, $1 < j < columns$). The matrix of the cell's contributions to the overall χ^2 value can now be easily calculated from Tables 6 and 7. This matrix is presented in Table 8.

Table 8: Contributions to the overall χ^2 value of the distribution of NPTYPE relative to SVORDER

	V1 Constructions	*S1 Constructions*
Lexical NP	$\frac{(356-284.26)^2}{284.26} \approx 18.11$	$\frac{(75-146.74)^2}{146.74} \approx 35.07$
Proper Name	$\frac{(14-25.72)^2}{25.72} \approx 5.34$	$\frac{(25-13.28)^2}{13.28} \approx 10.35$
Pronoun	$\frac{(0-60.02)^2}{60.02} \approx 60.02$	$\frac{(91-30.98)^2}{30.98} \approx 116.26$

Since under the null hypothesis each cell in the table is distributed $\chi^2(1)$ we can calculate the p value for each cell. However, since we are conducting six χ^2 tests, our probability of a type I error increases[42] and we should correct our p values. For this purpose I will be using the conservative bonferroni correction which essentially multiplies the p value by the number of tests—in our case 6. The corrected p values are displayed in Table 9.

[41] Cramer V is a post-hoc test that determines the strength of an association after a χ^2 test has determined its significance. The formula to calculate cramer's V is $V = \sqrt{\frac{\chi^2}{n(k-1)}}$ where χ^2 is the statistic obtained by the χ^2 test, n is the total number of table elements (i.e. total number of sentences) and k is the minimum between the number of rows and columns in the table. Cramer's V varies between 0 and 1. As a rule of thumb, value above 0.3 indicate strong correlation, values between 0.1 and 0.3 indicate intermediate correlation and values below 0.1 do not indicate correlation at all.

[42] Type I error, also known as a false positive error, is the claim that a non-significant result is significant. If one conducts six tests and in each the probability of error is p=0.05 then the total probability of making at least one error is greater than 0.05 and is equal to $1 - 0.95^6 = 0.265$.

Table 9: p-values for the χ^2 contributions corrected with the bonferroni corrections

	V1 Constructions	*S1 Constructions*
Lexical NP	p<0.001	p<0.001
Proper Name	p=0.008	p=0.12 (p=0.02 with Holm's correction)
Pronoun	p<0.001	p<0.001

As can be seen in Table 9, all cells vary significantly from their expected frequencies, except for the count of proper names in S1 sentences. The observed value of S1 sentences with proper names is indeed larger then the expected count of 13, but this difference is not statically significant. Note however, that to arrive at the significance score of p=0.12 for that cell we have used the conservative bonferroni correction. Use of the less conservative Holm correction yields a significant p-value (p=0.02). It was not necessary to employ the Holm correction for any other post-hoc tests in this work, but because of the obvious alignment of proper names with the S1 word order (the smaller than expected number of proper names in V1 constructions was highly significant even when using the bonferroni correction) it was employed in this one instance.

Following the above discussion and from the data in tables 6, 7 and 9, we can conclude that for pronouns and proper names the S1 word order is significantly more frequent than is expected by chance, whereas the V1 word order is significantly less frequent than expected by chance. Calculating the correlation coefficient for the two levels yields V=.61 for the *Pronoun* level and V=.17 for the *Proper Name* level[43]. for lexical NPs we get the opposite result, The V1 word order is significantly more frequent than expected by chance and the S1 word order is significantly less frequent than expected by chance (V=.64 for the *Lexical NP* level). The strong alignment of pronoun subjects with S1 sentences, weaker alignment of proper name subjects with S1 sentences and the strong alignment of lexical subjects with V1 sentences perfectly match our predictions for this factor in Table 5—*Pronoun*>*Proper Name*>*Lexical NP*.

Despite the evident correlations between NPTYPE and SVORDER, one may still ask if these correlations indicate that NPType affects SVOrder, or maybe these correlations are just epiphenomenal to the effects of other factors. For instance, as shown later in this section , the accessibility of the subject is also strongly correlated with word order where highly accessible subjects are aligned with the S1 order and low accessibility subjects are aligned with the V1 order. As was demonstrated by Ariel (1988, 1990, 2001) accessibility is tightly connected with the form of the NP. Highly accessible entities tend to be coded by pronouns, while non accessible entities tend to be coded by lexical NPs. Can it be the case that the strong correlations between pronouns and the S1 order and between lexical NPs and the V1 order are just a result of their levels of accessibility? In addition, the weaker but still significant correlation between proper names and S1 sentences can be argued to stem from the fact that proper names (in my data, mainly names of individuals) are typically animate. As shown later in this section,

[43] In order to calculate the correlation coefficient for a certain level, we create a contingency for the level and SVORDER. The new table has one row for the V1 and S1 counts of the specified level, and one row for the V1 and S1 counts of all sentences not in this level. We then calculate cramer's V for this new table.

animate subjects strongly prefer the S1 word order. Bearing this in mind, we can now also wonder if the association of proper names with S1 sentences is just an epiphenomenon of animacy[44].

The above discussion poses a question about the significance of the effect of NPType in the presence of other factors. Specifically, it argues that the effect of NPType might be reduced to Access and Animacy. These types of questions are relevant to many of the discussed factors and they are very hard to intuitively answer. In section 5.3 I will use statistical techniques (namely multifactorial regression) to bear on these issues. The analysis discussed in that section will provide a negative statistical answer to the above question regarding the factor NPType—the factor NPType contributes significantly to the choice of word order even when the influence of all other factors is considered. In the special case of NPType however, we can also intuitively sense that its contribution is not reducible. Specifically, we can note the existence of a grammatical constraint against bare pronouns in V1 constructions, a constraint which as far as I know is uncontested. In the presence of a pronoun this constraint allows us to predict the word order with a degree of certainty that cannot be obtained with other discourse old or animate entities. This situation nicely demonstrates one of the main arguments of this work: while diachronically, it is quite possible that the distribution of the different NP types over the V1 and S1 constructions was directed by factors such as accessibility and animacy, it seems that gradually, the mind identified the patterns of NPType and grammaticized them. In this case the grammaticization is strong and obvious so it can be noticed without the need for complex statistical procedures; in other cases it is more subtle and the aid of statistical procedures is required in order to decide whether a factor is reducible to others or not.

Definiteness The overall distribution of Def relative to SVOrder is highly significant ($\chi^2(1) = 102.08$; $p < 0.001$; $V = 0.43$). The frequencies in all cells varied significantly from the expected frequencies ($p<0.001$ for all cells beside Definite*V1 which was at $p=0.002$). Specifically, for definite subjects the S1 word order is significantly more frequent than expected by chance whereas the V1 word order is significantly less frequent than expected by chance; for indefinite subjects the V1 word order is significantly more frequent than expected by chance whereas the S1 word order is significantly less frequent than expected by chance. The distribution is therefore in accord with our prediction from Table 5—*Definite* > *Indefinite*.

Table 10: Distribution of Def relative to SVOrder

	V1 Constructions	*S1 Constructions*	*Totals*
Definite	184 (≈51%)	178 (≈49%)	362
Indefinite	186 (≈93%)	13 (≈7%)	199

Another interesting thing to note about the above distribution is that half of the V1 subjects were definite (184 definite vs. 186 indefinite subjects). While this number is still significantly lower than

[44] It should be noted that proper names in my corpus—as opposed to what one might think due to their non-pronominal coding—are not normally new entities. Only 10 out of 39 of the proper names in my sample were new (16 where old, 9 were assumed to be stored in the hearer's long term memory and 4 were inferable). The 10 *New* names however indicate that the significantly high frequency of proper names in S1 sentences cannot be reduced to accessibility and that it is better to explain it through Animacy (35 out of the 39 proper names were animate).

expected by chance (since there are overall more definite subjects than indefinite subjects), it clearly contradicts previous claims about a syntactic constraint against definite subject in V1 sentences (see section 2.1.4).

The distribution can be readily explained if we allude to topicality. As is well known, indefinite topical subjects are very rare. For an indefinite to be the topic it has to be either generic or a highly specific NP (cf. Giora 1981, p. 271-273; Erteschik-Shir 2007; inter alia). Since the prototypical sentence coded in the S1 word order has a topical subject it becomes obvious that indefinites will be rare in this word order. The distribution is than explained not by a hard constraint against definite subjects in the V1 word order, but rather by a soft constraint against indefinites in the S1 order.

Person The overall distribution of SVOrder relative to Person is highly significant ($p<0.001$ Fisher's Exact Test[45]). Specifically, for 1st person NPs we see that the S1 word order is significantly more frequent than expected by chance whereas the V1 word order is significantly less frequent than expected by chance (V=.43); for 3rd person we observe the opposite, the V1 word order is significantly more frequent than expected by chance whereas the S1 word order is significantly less frequent than expected by chance (V=.45). Due to the small sample size, cramer's V cannot be reliably obtained for the 2nd person level, but the observed frequencies for this level do vary significantly from the expected frequencies ($p=0.01$ on Fisher's Exact Test when contrasting the second row and the sum of the first and third columns). Specifically, 2nd person NPs are more frequent than expected in S1 sentences and less frequent than expected in V1 sentences.

Table 11 summarizes the data.

Table 11: Distribution of PERSON relative to SVORDER

	V1 Constructions	*S1 Constructions*	*Totals*
1st Person	1 (≈2%)	51 (≈98%)	52
2nd Person	0 (0%)	4 (100%)	4
3rd Person	369 (≈73%)	136 (≈27%)	505

The prediction for the factor PERSON in Table 5 was *1>2>3*. While we don't have enough data to decide if the tendency of 1st person toward the S1 group is stronger than that of the 2nd person, we can definitely conclude that the data do not contradict our hypothesis and that the data are in accord with the general direction of the hypothesis. It reflects a relation of *1,2>3*.

Case and Agreement It is well known that subject–verb agreement and the assignment of nominative case to the subject are much more stable in S1 than in V1 sentences (cf. Ziv, 1976, Preminger, 2009). This pattern starts to reveal itself in my data—subject–verb agreement is never broken in the S1 sentences but is broken 6 times in the V1 sentences (which is 1.6% of the total number of V1

[45] When more than 80% of the expected frequencies are below 5 (in this case offending cells are those of 2nd person), the χ^2 results become unreliable. For this reason I will occasionally use Fisher's Exact Test to arrive at the significance of the overall distribution.

sentences)—although the difference is not statistically significant (p=.19 Fisher's exact test). Furthermore, all sentences in my corpus had nominative subjects so the factor CASE was obviously insignificant (p=1 Fisher's exact test). The corpus data is summarized in table 12 below.

Table 12: Distribution of AGR and CASE relative to SVORDER

	V1 Constructions	*S1 Constructions*	*Totals*
+Agr	364 (≈66%)	191 (≈34%)	555
-Agr	6 (100%)	0 (0%)	6
Nom	370 (≈66%)	191 (≈34%)	561
Acc	0	0	0

Although the phenomena of broken subject–verb agreement and subjects with accusative case are well known to occur in V1 sentences, it is also known that they are rare, especially in written texts. Therefore, the failure to obtain a significant effect should be regarded a consequence of data sparsity rather than the lack of correlation. Due to this problem and in lack of a better alternative I will briefly exemplify and discuss this phenomena from a qualitative standpoint.

In section 3.2 I have argued that the function of the grammatical subject is to uniformly code aspects of propositions—such as topichood and agentivity—that typically appear together. Subjects then, to a degree, code topics. In this regard the phenomena of broken agreement and subjects with accusative case are no different than inverted word order. All these phenomena have the function of removing subject features from a non-topical or an otherwise defective subject. In Hebrew, the SV word order is the first and most frequent subject marker to be dispensed with, but the stronger markers—case and agreement—will sometimes be dispensed with as well[46].

To demonstrate these phenomena, observe the examples in (38):

(38) a. ve-az, ze pašut kara. hem(3P) yašnu(3P) yaxad, [...] hem ra'u(3P) seret,
and-then, it just happened. they slept together, [...] they saw a-movie,
yašnu(3P) mexubakim, hitnašku(3P) [...] hitxabku(3P)[47].
slept hugging, kissed [...], hugged.
'And then, it just happened. They slept together, saw a movie, slept hugging each other, kissed, hugged.'

b. ani lo yexola lišon, koev(3SM) li ha-beten(3SF), ani lo yexola yoter.[48]
I NEG able to-sleep, hurt to-me the-stomach, I NEG able anymore.
'I can't sleep, my stomach hurts, I can't do it anymore.'

[46] See Ziv (1976) and appendix B.2 for an analysis of this phenomenon with regard to the existence/possession predicate *yeš*. Based on my corpus data and other data I have worked with, it appears that the subject–verb agreement and the nominative case features are only lost when the sentence is in the VS order (i.e. when the word order indicator of the subject is lost as well). This is to my knowledge a hard constraint. It also appears to be that with regard to verbs other than *yeš* and its inflections, agreement is less stable than case. That is, the assignment of accusative case to the subject is a more rare condition than the lose of agreement. Again, at this point this is an observation and not a statistical conclusion. I only point it out here since it is contrary to Ziv's data about the existence/possession predicate and to Kenaan's promotional hierarchy (see discussion in appendix B.2) and it may be an interesting issue for future research.

[47] Blog entry: http://www.tapuz.co.il/blog/ViewEntry.asp?EntryId=780598

[48] Health forum: http://sc.tapuz.co.il/communa-3276-75-.htm

c. wa'i, kara li et hadavar haxi muzar b-a-olam.[49]
wow, happened to-me ACC the-thing most strange in-the-world.
'Wow, the strangest thing in the world happened to me.'

d. yeš et ha-sfarim(3PM) šel ha-'universita ha-ptuxa šel ha-kurs mivnim
EXIST ACC the-books of the-university the-open of the-course structures
algebriyim, sax ha-kol tovim.[50]
algebraic, all-in-all good.
'The Open University has books for the course "algebraic structures". These books are overall quite good.'

In (38-a) the subject-topic of the clauses are a pair of lovers coded by the pronoun *hem* 'they'. In all these clauses the subject—the lovers—are strongly perceived as the topic, so all subject markers are kept—word order, case and agreement.

In (38-b), in the clause *koev li ha-beten* 'hurts to-me the-stomach', the speaker is topical and the new information about her is that her stomach aches. However, in this case the speaker is not coded as the subject and as a result, the subject is not topical. The V1 word order is then selected (the subject's word order marking is lost), but also subject–verb agreement is broken.

Sentence (38-c) is similar. The speaker is topical, but she is not the subject. The V1 order is again selected, but this time subject–verb agreement is preserved and it is the nominative case feature that is dispensed with—the subject *hadavar haxi muzar b-a-olam* 'the strangest thing in the world' appears in the accusative case.

In Sentence (38-d) all three subject markers are lost: the sentence is V1, subject-verb agreement is broken, and the accusative case is assigned instead of the nominative. This is the unmarked behavior of the existence/possession predicate *yeš* and it is to a large extent grammaticized (cf. Ziv, 1976)[51]. This finding is not surprising. The existence predicate normally appears with non-topical subjects so it makes sense for it to appear in V1 constructions. However, the existence predicate is also by far the most frequent predicate in these constructions (and probably the most frequent predicate in Hebrew) so it is reasonable that grammatical features that are associated with V1 constructions will be more entrenched in its case.

It should be stressed that I do not maintain that marking of non-topical subjects can account for all the intricacies of phenomena such as non-nominative subjects and broken agreements. Indeed, to account for the loss of agreement, Preminger (2009) suggests that syntactic structure is a determining factor; other factors such as the subject's animacy or additional sentence elements with different ϕ-features are probably at work as well[52]. The fact remains that a discrepancy between S1 and V1 sentences with

[49] Soccer forum: http://www.asoccer.co.il/index.php?showtopic=18407&st=380

[50] Computer science forum: http://sf.tapuz.co.il/shirshur-1428-69364251.htm

[51] Excluding extreme cases of contrastive focus the word order of the existence predicate *yeš* 'there-is' is fixed on V1. The predicate can occasionally agree with its subject in gender and number, but this paradigm (i.e. *yešno*, *yešna*, *yešnam* etc.) is very uncommon in colloquial Hebrew and is the sign of a high register. The past and future inflections of this predicate (*haya* 'there-was' and *yihiye* 'there-will-be') still occasionally agree with their subject in Spoken Hebrew, but as argued by Ziv (1976), this too may be subject to change in the future.

[52] My attitude towards modeling the data of breaking agreement would also involve multifactorial models. Indeed, case and agreement are themselves subject markers; it thus makes sense to treat them as output variables and model them using many of the factors discussed here. Preminger (2009) favors a syntactic approach, but I believe his basic assumptions are falsified by corpus data. A basic assumption of his model is that agreement always holds in [V S] configurations, that is, in V1 sentences without an intervening element between the verb and the subject. While this might be true as a statistical tendency (I have no data to bear on this issue), it is definitely not true as a categorical restriction. *eyzo aruxa dafakti. koev(3SM) ha-beten(3SF)* 'what a meal I've had. hurts the-stomach' (url: http://www.shin1.co.il/ya.php?sid=1701493) or *niš'ara(3SF) sug šel 'ironia(3MF)* 'remained a-type of irony' from the Linzen's blog corpus are two examples of such sentences, but there are of course others.

regard to the subject–verb agreement and the subject's case features is expected from the association of these features with low topicality and that this expectation is indeed born out.

5.2.2 Semantic Factors

Animacy The overall distribution of ANIMACY relative to SVORDER is highly significant $\chi^2(1) = 188.58$; $p < 0.001$; $V = 0.58$. The frequencies in all cells varied significantly from the expected frequencies ($p<0.001$ for all cells). Specifically, for animate subjects the S1 word order is significantly more frequent than expected by chance whereas the V1 word order is significantly less frequent then expected by chance; for inanimate subjects the V1 word order is significantly more frequent than expected by chance whereas the S1 word order is significantly less frequent than expected by chance. The distribution is therefore in full accord with our prediction from Table 5—*Animate* > *Inanimate*.

Table 13 summarizes the data.

Table 13: Distribution of ANIMACY relative to SVORDER

	V1 Constructions	*S1 Constructions*	*Totals*
Animate	64 (≈30%)	147 (≈70%)	211
Inanimate	306 (≈87%)	44 (≈13%)	350

Agentivity The overall distribution of AGENTIVITY relative to SVORDER is highly significant $\chi^2(1) = 102.23$; $p < 0.001$; $V = 0.43$. The frequencies in all cells varied significantly from the expected frequencies ($p<0.001$ for all cells beside Non-Agentive*V1 where $p=0.009$). Specifically, for agentive subjects the S1 word order is significantly more frequent than expected by chance whereas the V1 word order is significantly less frequent then expected by chance; for non-agentive subjects the V1 word order is significantly more frequent than expected by chance whereas the S1 word order is significantly less frequent than expected by chance. The distribution is therefore in full accord with our prediction from Table 5—*Agentive* > *Non Agentive*.

Table 14 summarizes the data.

Table 14: Distribution of AGENTIVITY relative to SVORDER

	V1 Constructions	*S1 Constructions*	*Totals*
Agentive	46 (≈32%)	100 (≈68%)	146
Not Agentive	324 (≈78%)	91 (≈22%)	415

Verb Class The overall distribution of VCLASS relative to SVORDER is highly significant $\chi^2(2) = 190.27$; $p < 0.001$; $V = .58$. Post hoc tests reveal however , that while the contribution of the *Unaccusative* and *Unergative* levels to the overall χ value was highly significant ($p<0.001$ for all relevant cells), the contribution of the *Passive* level was negligible ($p=1$). The analysis reveals that for

unaccusative verbs the V1 word order is significantly more frequent than expected by chance whereas the S1 word order is significantly less frequent then expected by chance (V=.55); for Unergative verbs the S1 word order is significantly more frequent than expected by chance whereas the V1 word order is significantly less frequent than expected by chance (V=.56). The correlation of the *Passive* level with SVOrder was as expected very low (V=.02).

Table 15 summarizes the data.

Table 15: Distribution of VClass relative to SVOrder

	V1 Constructions	*S1 Constructions*	*Totals*
Passive	21 (≈62%)	13 (≈38%)	34
Unaccusative	322 (≈83%)	64 (≈17%)	386
Unergative	27 (≈19%)	114 (≈81%)	141

In the literature unaccusatives and passives are usually lumped together and are both argued to sound natural in the V1 word order (cf. Reinhart and Siloni, 2004b, Shlonsky, 1987, 1997). For this reason our prediction for the factor VClass in Table 5 was *Unergative* > *Unaccusative*, *Passive*. From my data it appears the correct scale is *Unergative* > *Passive* > *Unaccusative*. It should be stressed however that the literature referenced above did not make any claims about the relative ordering of passives and unaccusatives with regard to word order, the only claim was that that both levels sound more natural in V1 constructions then the group of unergative verbs. In that respect my results are in accord with the predictions, and in fact, further specify them.

5.2.3 Discourse Pragmatic Factors

Accessibility The overall distribution of Access relative to SVOrder is highly significant $p < .001$ *Fisher's Exact Test*; $\chi^2(3) = 212.20$, $p < .001$; $V = .62$. The frequencies in the cells of the *New* and *Old* levels varied significantly from their expected frequencies (all p<0.001). Specifically, for *Old* NPs we see that the S1 word order is significantly more frequent than expected by chance whereas the V1 word order is significantly less frequent than expected by chance (V=.58); for *New* NPs we observe the opposite, the V1 word order is significantly more frequent than expected by chance whereas the S1 word order is significantly less frequent than expected by chance (V=.58).

Further analysis reveals that the highly significant overall distribution results entirely from the *New* and *Old* levels. The levels *Inferable* and *LTM* did not vary significantly from their expected frequencies[53] and the correlation coefficients for these levels also do not indicate any correlation (V=.03 for Inferables and V=.08 for LTM).

Table 16 summarizes the data:

[53] Post-hoc χ^2 for the cells of both levels followed by the bonferroni correction gives the value of p=1 for both. However, in the case of the *LTM* level, because of the low number of *LTM* subjects, the p-value for this level is not reliable. For this reason I have conducted Fisher's exact test contrasting the second row with the sum of the three others (i.e. *LTM* vs. *non*-L*TM* subjects). The results were insignificant (p=.09) even before correcting the p-value with bonferroni. I conclude that the levels of Inferable and *LTM* do not change significantly from their expected frequencies.

Table 16: Distribution of ACCESS relative to SVORDER

	V1 Constructions	*S1 Constructions*	*Totals*
Inferable	28 (≈61%)	18 (≈39%)	46
LTM	6 (≈43%)	8 (≈57%)	14
Old	45 (≈26%)	131 (≈74%)	176
New	291 (≈90%)	34 (≈10%)	325

The strong tendency of sentences with *Old* subjects toward the S1 word order together with the strong tendency of sentences with *New* subjects toward the V1 word order and the intermediate behavior of sentences withe *LTM* or *Inferable* subjects (that did not reveal an above chance preference for any of the word orders) are in perfect accord with our predictions in table 5—*Old* > *Inferable*, *LTM* > *New*.

5.2.4 Other Factors

NP Length The average length of subjects in S1 sentences is 1.36 words (SD=0.80); the average length of subjects in V1 sentences is 2.79 (SD=2.31). The difference is highly significant ($t_{welch}(508.68) = 10.73$; $p < 0.001$; $r_{pb} = 0.33$). Another way to look at the difference in word lengths between the V1 and S1 sentences is to cross tabulate LENGTH and SVORDER for the different word lengths while maintaining a single cell for subjects whose length is greater than a certain threshold (this technique was demonstrated in Gries 2003). The product of this cross tabulation is presented in Table 17 where we clearly see that subjects of length 1 are much more common in the S1 order then expected by chance (remember that because of the larger proportion of V1 sentences in our corpus, our expectation under the null hypothesis is that only around 34% of the sentences with 1 word subjects would be S1, but the observed results report around 54%). But for subjects of lengths greater than 1, the results turn and they appear in the V1 word order more often than expected by chance (and this preference for the V1 order becomes stronger for longer subjects). All the data is then in accord with our prediction that the S1 sentences will have shorter subjects than V1 sentences.

Table 17: Distribution of LENGTH relative to SVORDER

	V1 Constructions	*S1 Constructions*	*Totals*
1	124 (≈46%)	147 (≈54%)	271
2	106 (≈78%)	30 (≈22%)	136
3	50 (≈86%)	8 (≈14%)	58
4	30 (≈91%)	3 (≈9%)	33
5	20 (≈91%)	2 (≈9%)	22
6	13 (≈93%)	1 (≈7%)	14
≥7	27 (100%)	0 (0%)	27

5.2.5 Summary and Conclusions

Summarizing the discussion in the previous section, it appears that 8 out of the 10 discussed factors provided statistical results that fully complied with our predictions. Furthermore the data for the two remaining factors—AGR and CASE—did not contradict our predictions, but were simply too sparse to provide meaningful insight. Qualitative data were thus used to discuss these factors and to argue that they too fall into the expected pattern.

The results of the monofactorial analysis are summarized in table 18. The factors are ranked by their correlation coefficients that indicate the strength of the correlation. It should be noted however, that the coefficient for the factor Length is not directly comparable to the other coefficients since Length is measured in a different scale than the other factors. The correlation coefficients are provided for each factor as a whole. For the coefficients of the individual levels the reader is referred to the detailed analysis in sections 5.2.1–5.2.4.

Table 18: Predicted vs. observed results for the different factors

Factor	*Correlation*	*Predicted Behavior*	*Observed Behavior*
NPType	V=0.66	*Pronoun* > *Proper Name* > *Lexical NP*	*Pronoun* > *Proper Name* > *Lexical NP*
Access	V=0.62	*Old* > *Inferable, LTM* > *New*	*Old* > *Inferable, LTM* > *New*
Animacy	V=0.58	*1* > *0*	*1* > *0*
VClass	V=0.58	*Unergative* > *Passive, Unaccusative*	*Unergative* > *Passive* > *Unaccusative*
Person	V=0.45	*1* > *2* > *3*	*1,2* > *3*
Agentivity	V=0.43	*1* > *0*	*1* > *0*
Def	V=0.43	*1* > *0*	*1* > *0*
Length	$r_{pb} = 0.33$	Len(S1 Subjects) < Len(V1 Subjects)	Len(S1 Subjects) < Len(V1 Subjects)
Agr	V=.07	*1* > *0*	—
Case	-	*Nominative* > *Accusative*	—

The results of this section provide ample evidence for the first part of the topicality hypothesis—(36-a). All factors levels that are known to be associated with high topicality appeared more often than chance in S1 sentences and all factors levels that are known to be associated with low topicality appeared more often than chance in V1 sentences. Despite the large number of factor and factor levels discussed, there were no exceptions to the expected behavior.

Despite these encouraging results, one has to be careful in concluding that all of the above factors are active in the choice of word order. At this point it has not been determined whether the correlations of the different factors constitute an independent contribution to the choice of word order or are just epiphenomenal to the effects of other factors. Determining which of the factors contribute independently to the choice of word order is the subject of the next section.

5.3 Multifactorial Results

5.3.1 Classification Tree

To open our discussion of multifactorial models, I have fitted the data with a classification tree model containing all eight relevant factors from section 5.2. The formula for this initial model is given in (39) below:

(39) Classification Tree Model
SVOrder ~ VClass + NPType + Access + Length + Def + Animacy + Person + Agentivity

The formula indicates that the model should attempt to probabilistically predict the value of the dependent variable SVOrder based on the values of the eight listed factors. Figure 4 shows the tree for the model.

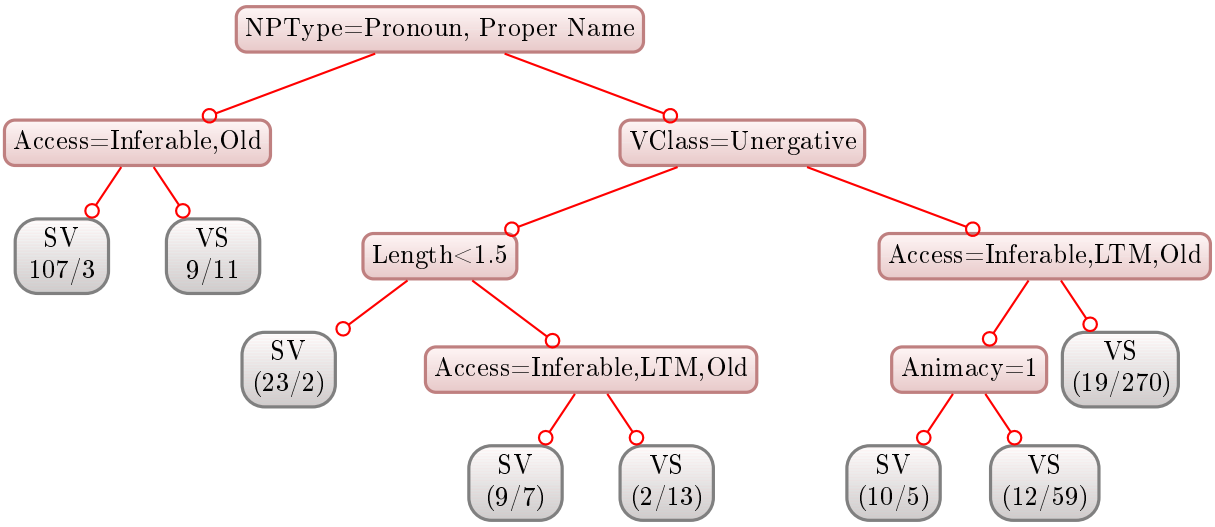

Figure 4: Graphical outline of the Classification Tree Model

The tree outlines a procedure to determine the sentence word order based on its features. Each node in the tree presents a condition to determine the next node to visit. We start at the root node and examine its condition. If the condition is met we continue to the left node, otherwise we continue to the right. This procedure ends once we reach a leaf node. Leaf nodes are labeled either *SV* or *VS* and determine the predicted word order for the sentence. For example, if our input sentence has a pronoun or a proper name subject we go left, then if the subject is also inferable or old we go left again and arrive at a terminating leaf node predicting that the sentence is in the SV order. This prediction is supported by 107 observations and contradicted by 3 so we can also conclude that under these conditions the probability for the SV order is 107/110≈97% and the probability of the VS order is approximately 3%.

The procedure for constructing the tree is also straightforward. The algorithm first inspects all predictors and chooses the most useful one to populate the root node. The algorithm then continues to grow the tree in a similar manner: for each new node to be created, the algorithm inspects all predictors (including the ones already used) and chooses the one most useful for this stage of the analysis[54]. The algorithm stops once a node holds less than 20 observations or once creating new nodes is not likely to improve performance.

The classification tree provides a convenient way to understand the trends in the data and to get a feel for some existing interactions. For instance VCLASS appears only in the right branch of the tree, so according to the tree model it is only relevant for clauses whose subject is a lexical NP. We have then an interaction of VCLASS and NPTYPE. The same holds for LENGTH that interacts with NPTYPE but also interacts with VCLASS. ANIMACY appears to be relevant only for a combination of a *Lexical NP* subject, an *Unaccusative* or *Passive* verb and an *Old*, *LTM* or *Inferable* Subject, and so it interacts with NPTYPE, VCLASS and Access. In this way the tree model can clearly outline high order interactions.

[54] There are different algorithms to decide what is the most useful factor in any given stage, but the most basic and straightforward one simply chooses the factor that splits the remaining sentences in the most extreme manner with regard to word order, and thus provides for the maximal information gain.

The performance of the tree model is quite good and when trained and tested on our training corpus it accurately predicts 89.66% of the data. To test for overfitting, the model accuracy scores were averaged over 200 bootstrapping iterations[55] (resulting in accuracy of 88.72%) and averaged again over 561 iterations of leave one out cross validation[56] (resulting in accuracy 86.27%). These initial results indicate a degree of overfitting, but even so, they are encouragingly high considering that the baseline for this problem—a model that always predicts the V1 word order—only arrives at an accuracy score of 65.95% .

The classification tree model has an advantage over the regression models I will soon discuss, in that it picks up on interactions between factors by itself , and outlines them in a way that is easily interpreted by the linguist. This model has however a few downsides: firstly, at least as far as the open source R software package is concerned, classification trees do not provide significance levels and p-values for the different factors. While there are ways to arrive at these values, the manual work will be unnecessarily tedious. Secondly, classification trees provide a mechanism according to which the speaker sequentially considers the values of the different factors. This model was suggested by Gries (2003, p. 115,116) to be cognitively less plausible than that of regression models in which all factors are considered simultaneously.

Based on these arguments I will not develop the tree model further. It is presented here for illustrative purposes and because it can point out interactions that one can later entertain in the interaction modeling stages of logistic regression. For a more in depth discussion of the procedures discussed here, as well as for the ways to tune and validate classification tree models, the reader is referred to Baayen (2008, p. 148-154)

5.3.2 Logistic Regression

Preliminaries and mathematical formulation Logistic regression is probably the most common way to model a binary response variable whose value is influenced by both ordinal and continuous factors. This modeling technique is becoming prevalent in linguistics following the works of Williams (1994), Arnold et al. (2000), Bresnan et al. (2007) and also Gries (2003) who used the related technique of discriminant analysis. The predicted probability of the binary variable SVOrder under the assumptions of the logistic model is given in formula 1.

$$P(SVOrder = \text{"}VS\text{"}) = \frac{1}{1 + e^{-\beta X}} \tag{1}$$

Where $X = (X_0, X_1, ..., X_n)$ is a vector of variables corresponding to the different levels of our factors, and $\beta = (\beta_0, \beta_1, ..., \beta_n)$ is a vector of coefficients. Each coefficient β_i describes the size of

[55] Bootstrapping is a validation procedure in which we randomly sample from our training corpus a group of sentences that's the size of corpus, but we do it with replacement. For instance since our training corpus contains 561 sentences, taking from it a sample of 561 sentences with replacement will result at a new training set of 561 sentences, of which there are around 350 unique sentences (the rest are double instances of these 350). We then train our model on this new corpus and test it against the original corpus. In this way we both have a reasonably large training corpus, and a large test corpus that contains a large group of sentences the model has never seen before (≈561-350=211). Averaging over a large number of bootstrapping iterations should provide us with information about our model's performance on unseen data and cancel out the effects of overfitting.

[56] Leave one out cross validation is another validation technique designed to account for overfitting. In leave one out cross validation we train the model on all the data except for a single sample and then test the model on that one sample. We repeat this process for each sample in our corpus and average over the obtained results. In each iteration we have a large training corpus, yet our model predicts the word order of an unseen sentence.

the contribution of its corresponding factor level x_i. Before starting the regression process all nominal factors are converted to binary variables and mapped to the variables X_0, X_1, ..., X_n[57]. The goal of the modeling process is to match the variables $X = (X_0, X_1, ..., X_n)$ (i.e. our factor levels) with a vector $\beta = (\beta_0, \beta_1, ..., \beta_n)$ of coefficients that determines their relative strength.

As indicated by formula 1, weighing the factors X by their weights β (i.e. the product $X\beta = X_0\beta_0 + X_1\beta_1 + ... + X_n\beta_n$) does not directly provide the probability of the VS word order. However, the relation between Xβ and the probability is such, that the higher the value of $X\beta$ the higher the probability of the VS word order[58]. In particular, the coefficients β and the weighted sum $X\beta$ can be negative—indicating a preference for the SV order, or positive—indicating a preference for the VS order. When $X\beta = 0$ then the probability for VS is equal to the probability of SV (i.e. $P(\text{"}VS\text{"}) = 0.5$). To arrive at exact word order probabilities from the vectors X and β, we'll use formula 1.

Modeling Process and Initial Results Limitations of logistic regression necessitated the elimination of factor levels that are deterministic with regard to word order or that otherwise do not contain enough data points. Namely, the level Access=*LTM* was brought together with Access=*Inferable* (since the *Access=LTM* level had only 14 data points); The level NPType=*Pronoun* was brought together with NPType=*ProperName* (the level NPType=*Pronoun* is deterministic—all 91 subject pronouns appeared in the SV order) and the level Person=*2* was brought together with the level Person=*1* (The level Person=2 had only 4 data points, all in the SV order). The unification of levels with few data points is largely insignificant from the standpoint of model strength. However, the unification of the pronoun level with the proper name level unified two levels with an adequate number of data points and was conducted just because the Pronoun level had no V1 samples. This unification slightly reduces the predictive force of the NPType factor since the deterministic nature of the pronoun level is lost[59].

Following these preliminaries I defined model A to include all eight factors and fitted it to the data. The formula for the model is given in (40).

(40) Model A
SVOrder ~ VClass + NPType + Access + Length + Def + Animacy + Person + Agentivity

[57] The way in which this mapping is done is by equating the influence of one level of a factor with the intercept β_0—whose corresponding variable x_0 is defined to be 1—and then creating a variable for each additional level of the factor. In this way the effect of the levels mapped to the intercept is always present, and the effects of other levels are added to it if the respective variables are given the value 1. For example, the factor VClass that has the levels Unaccusative, Unergative and Passive can be coded as two binary variables lets say x_1 and x_2 that map to the levels VClass=Unaccusative and VClass=Unergative. The level VClass=Passive is mapped to the intercept and affects the value of β_0. If both x_1 and x_2 are 0 (i.e. the sentence involves a passive verb) than the effect of the level passive will bear on the outcome probability through the intercept; if however one of the variables x_1 or x_2 is 1 (i.e. the verb is either unaccusative or unergative), then its respective β value (β_1 or β_2) will be taken into account so as to override the probability assigned by the intercept alone.

[58] It should be stressed however, that the relation between $X\beta$ and the probability is not linear. It is mediated by the logit function—$(*)$ $X\beta = logit(P(SVOrder = \text{"}VS\text{"}))$. The logit function is given by $logit(x) = log(\frac{x}{1-x})$; its inverse—the logistic function—is given by $\pi(x) = \frac{e^x}{1+e^x}$. We can arrive at the probability given in formula 1 by applying the logistic function to both sides of the equation $(*)$ as follows: $\pi(X\beta) = \pi(logit(P(SVOrder = \text{"}VS\text{"})) \Rightarrow \frac{e^{X\beta}}{1+e^{X\beta}} = P(SVOrder = \text{"}VS\text{"}) \Rightarrow P(SVOrder = \text{"}VS\text{"}) = \frac{1}{\frac{1+e^{X\beta}}{e^{X\beta}}} \Rightarrow P(SVOrder = \text{"}VS\text{"}) = \frac{1}{1+e^{-X\beta}}$.

[59] This might have been a problem if the factor NPType would have turned out insignificant, since in that case we would not be able to know its significance level had the regression model did not have the said limitation. Luckily, this does not apply here as the factor NPType turns out to be significant under all models considered.

The model's likelihood ratio (its deviance relative to the deviance of the null model)[60] is 398.5 (df=10) which is highly significant ($p<0.001$). Its concordance score (C=0.934) is also very high which indicates excellent discrimination between the SV and VS levels[61]. The accuracy scores for the model—summarized in table 19—are 88.95% (i.e. the model correctly predicts word order in 88.95% of the corpus sentences).

Table 19: Accuracy scores for Model A (cutoff=50%)

	Predicted="SV"	Predicted="VS"	Precision
Observed="SV"	149	42	74.35%
Observed="VS"	20	350	94.59%
Overall			88.95%

To check for overfitting effects on performance I have averaged the model accuracy results over 200 bootstrapping iterations (88.88%), and over a full cycle of leave one out cross validation (87.34%). Also, comparison of the concordance score C between the training and test sets of 200 bootstrapping iterations yields an optimism level of (.0058) and a corrected concordance score of C=0.929[62]. After correcting for overfitting, these results are already better than the corrected results of the classification tree model presented earlier.

To arrive at the set of significant factors, I have used the likelihood ratio test (the statistic D). For each factor in model A, D is the difference between the model's deviance (321.08) and the deviance of the model without that factor. Under the null hypothesis this difference is distributed χ^2 with degrees of freedom equal to the number of degrees of freedom of the removed factor. A significant value of D indicates that the factor is required in the model even in the presence of all other factors. Table 20 indicates that the significant factors are VClass, Length, Access, NPType, Def and Animacy (Animacy is marginally significant). All these factors thus effect the choice of word order in a way that is not reducible to the effect of other factors.

Interestingly the factors Agentivity and Person did not turn out significant once the influence of other factors was considered. In the case of Agentivity this was to be expected since it is highly correlated with both VClass (V=0.64 for their correlation) and Animacy (V=0.71). It appears that the other semantic factors do a better job in predicting word order than Agentivity and in effect render it insignificant (indeed removing Animacy and VClass from the model results in an inferior model, but one in which Agentivity is highly significant). In the case of Person, it is highly correlated only with NPType (V=0.61), but it seems this correlation is sufficient to render it

[60] The deviance of the model is related to the likelihood the model ascribes to its training corpus (mathematically its -2 the difference between the log-likelihood of the model and the log-likelihood of a saturated model). The higher this probability the lower the deviance. The deviance in logistic regression plays the role of the residual error in linear regression (i.e. a measurement of the variance not explained by the model).

[61] The C index—the probability of concordance between predicted probability and response—is derived from the Wilcox-Mann-Whitney two sample rank test. It is computed by taking all possible pairs of sentences such that one of the sentences is SV ordered and the other VS ordered; The index is the proportion of these pairs and the subset of pairs in which the VS sentence was indeed equated by the model with a higher probability then the SV one (cf. Harrell, 2001, p. 247).

[62] The optimism level is the difference in C score between the training and test sets averaged over 200 bootstrapping iterations. This score is then subtracted from the original C score to arrive at the C score that is corrected for overfitting.

Table 20: The likelihood ratio D for Model A with a single factor excluded

Excluded Factor	*Df*	*D*	*P*
VClass	2	31.91	p<.001
Length	1	24.78	p<.001
Access	2	17.03	p<.001
NPType	1	8.73	p=.003
Def	1	6.93	p=.008
Animacy	1	4.32	p=.038
Person	1	1.32	p=.251
Agentivity	1	0.15	p=.697

insignificant. Basically all the 1st and 2nd person NPs in my corpus were pronouns, so with regard to these levels PERSON has no predictive advantage over NPTYPE (since sentences including pronouns are already predicted by NPTYPE to be S1). However, for 3rd person NPs, NPTYPE is much better in predicting word order since it differentiates between the pronoun, proper name and lexical NP levels, which all behave differently with regard to word order.

In order to further demonstrate the need for multiple factors I compared the model predictive power to the predictive power of models that rely on a single factor. The results in table 21 below are corrected for overfitting using both bootstrapping and leave one out cross validation (see notes 55 and 56 on page 51 for discussion of these techniques). It should be noted at this point that accuracy scores, despite their initial appeal, are generally frowned upon by statisticians as a measure of model strength. Harrell (2001) discusses their shortcomings at length and indeed the R statistical package does not provide them by default. The main problem with accuracy scores is that they fail to reflect the margin of the model's successful and unsuccessful predictions. Consider two models: the first model equates every SV sentence with 0% probability (probabilities are taken to be the probability of a *VS* outcome) and every VS sentence with 49.9% probability; the second model equates all sentences with 30% probability. We intuitively understand that the first model is much better than the second, but if the cutoff is taken to be 50%[63], both models will achieve precisely the same accuracy scores since for any given sentence both will always predict the word order to be SV. Beside the disregard of success/error margins, this example also demonstrates another important shortcoming of accuracy measurements—their sensitivity to the cutoff point. Indeed, if the cutoff point was just a little below

[63] the cutoff is the probability point that separates SV from VS predictions. Below it we consider the model's prediction to be SV and above it we consider the model's prediction to be SV. The cutoff for all accuracy scores I will present here is 50%, although sometimes the cutoff that achieves the best accuracy scores is not exactly on 50%, but rather a little below or above it.

Table 21: Accuracy scores for monofactorial models

Factor	*Bootstrapping Accuracy (n=200)*	*Leave One Out Cross validation Accuracy*	*Bootstrapped C (n=200)*
NPType	84.14%	84.14%	0.784
VClass	81.31%	81.46%	0.779
Access	81.06%	81.64%	0.83
Animacy	80.75%	80.75%	0.798
Agentivity	75.58%	75.58%	0.7
Person	75.58%	75.58%	0.643
Length	68.56%	70.05%	0.74
Def	65.54%	64.88%	0.718

the point of 50%, the first model would suddenly achieve perfect accuracy! In order to circumvent these problems I will always present the models' C scores in addition to accuracy scores, and indeed, the C scores should be taken to be more authoritative (for a discussion of C scores see footnote 61 on page 53).

As can be seen in table 21, our initial model (that does not yet take interactions into account) is already much stronger than the best single factor model. This is reflected by both the accuracy scores and the C scores[64].

The results presented verify the thesis that no single factor can account for the choice of word order and that it is multiple factors that determine that choice. According to our initial model—Model A—the factors VCLASS, LENGTH, ACCESS, NPTYPE, DEF and ANIMACY make significant contributions to word order and are not reducible to the effects of other factors. These results verify and support the second part of the topicality hypothesis (36-b).

Interactions In the previous section I fitted the data with an initial model—Model A—that considered all factors but did not consider interactions between them. From a linguistic standpoint, a slight shortcoming of logistic regression (when compared, for instance, to classification trees) is that it only adjusts the relative strength of the factors supplied to it—it does not uncover interactions by itself.

[64] One may initially think that the difference between an accuracy score of around 84% is not that far from an accuracy score of 88%, however that is definitely not the case. As long as the difference in accuracy persists over multiple bootstrapping or leave one out iterations—which it does—even a very small difference can be considered highly significant, let alone a difference of 4%. Note also that when accuracy approaches 90% every (fraction of) percent counts. Another way to look at such a difference is to note that the best monofactorial model makes around 30% more prediction errors then our initial multifactorial model.

The study of interactions in a model that contains eight highly correlated factors is a very complex process. In this section I do not intend to present a definitive study of the significant interactions in the data, but rather to show how outlining interactions can improve our model's performance and influence our understanding of the phenomenon.

As an initial step I have defined Model B to include the six significant factors from Model A, plus all their second order interactions[65]. The formula for model B is given in (41).

(41) Model B
SVOrder ~ (VClass + NPType + Access + Length + Def + Animacy)^2

This resulted in a much better fit for the data L.R=464.59 df=33 p<.001[66]. Calculating the difference of deviance between Model B and Model A reveals that Model B fits the data significantly better D=66.0841, df=23, p<.001 and the model also discriminates better (C=.958). However, a high proportion of this difference seems to be the result of overfitting—Model B seems to significantly overfit the data. When subject to bootstrapping, its discrimination score drops to C=.924—below that of model A—and its accuracy scores on bootstrapping (90.6%) and cross validation (89.1%) tests drop as well, although they still outscore those of model A. These results indicate a significant degree of overfitting, but they are not surprising. It is indeed expected that adding all possible interactions to the model will cause it to pick up on many patterns that are idiosyncratic to the specific training corpus.

The next step is then to take a minimal model that includes the six significant factors of Model A, adding to it only the interactions that appear significant. This stage required some trial and error, but it seems that adding the interactions between Animacy and Access and between Animacy and Def provides for the maximal gain in fit and performance with the minimal amount of clutter. The reason to this is clear once we examine the contingency tables for these interactions (outlined in table 22 below).

Table 22: Effects of Animacy over levels of Access and Def

	Access=*New*		Access=*Old*		Def=*0*	
	Animate	*Inanimate*	*Animate*	*Inanimate*	*Animate*	*Inanimate*
VS	44	247	9	36	30	156
SV	10	24	116	15	8	5

As the table indicates, *New* subjects tend to favor the VS order but this tendency varies depending on the levels of Animacy. For animate subjects this tendency is quite weak whereas for inanimate subjects it is very strong. An even more extreme case is that of *Old* subjects: sentences with subjects that are both *Old* and *Inanimate* are distributed according to expectations between the VS and SV levels (keep in mind that the number of VS sentences in the corpus is nearly twice the number of SV sentences) whereas sentences with subjects that are *Old* and *Animate* strongly prefer the SV word order. The same logic accounts for the Def*Animacy interaction as well. While in other instances the accumulative main effects of the factors can account for stronger tendencies toward one of the

[65] I excluded a single interaction—Def=*1* * NPType=*Pronoun,Proper_name*—because these values coincide completely and throw off the regression calculations at various stages of the analysis.
[66] The model's L.R score reflects the difference between its deviance and that of the null model.

two word orders, with regard to ACCESS, DEF and ANIMACY this is clearly not viable and thus the interaction effects are significant.

Following this discussion I created Model C and fitted it to the data. The formula for Model C is given in (42) where * marks an interaction between two factors:

(42) Model C
SVORDER ~ VCLASS + NPTYPE + ACCESS + LENGTH + ANIMACY + ANIMACY * ACCESS + ANIMACY=0 * DEF=0

Note that in this model, the main effect for the factor DEF is not considered. Model C only considers DEF's effect in the context of its interaction with ANIMACY, and more precisely, it only considers the interaction between indefinite and inanimate subjects (ANIMACY=0 * DEF=0). This is the result of trial and error in an effort to arrive at a parsimonious model that also achieves optimal results. It seems that when considering all other factors and interactions the main effect of DEF becomes only marginally significant and does not seem to reliably contribute to accuracy or discrimination scores. This of course does not diminish the significance of the contribution of DEF—if an interaction that includes the factor is significant the factor is obviously significant as well. Model C results are excellent and it definitely outperforms Model A. Model C's L.R score is 413.3 ($p<0.001$), its C score is 0.94 and accuracy score is 89.48%. When correcting for overfitting the C score is still very high C=0.935, and it achieves an accuracy score of 89.3% in leave one out cross validation and 89.86% in bootstrapping averaged over 200 iterations (the bootstrapping result is even better than the uncorrected accuracy scores, an indication that our model does not overfit the data). Model C's results are all significantly better than those of Model A, but as opposed to Model B, Model C does not overfit the data and as I will later show all of its factors and interactions are significant. Table 23 summarizes performance data for the three models.

Table 23: Performance comparison for the three regression models

	Df	Deviance	C	Accuracy	Bootstrapped C (n=200)	Leave One Out Accuracy	Bootstrap Accuracy (n=200)
Model A	10	321.082	0.934	88.95%	0.929	87.34%	88.88%
Model B	33	254.998	0.958	91.98%	0.924	089.1%	90.6%
Model C	10	306.282	0.94	89.48%	0.935	89.3%	89.86%

As can be seen in the table, both models B and C appear to be overall better than model A (although a case can be made for model A's strength compared to model B due to its improved C score after correcting for overfitting). When comparing Model B to Model C, I would conclude that Model C is stronger. Model B does fit the data significantly better (p=0.0006 for the difference between their respective deviances) but this is clearly the result of overfitting. When comparing accuracy results (cutoff=50%), Model B appears to be stronger on bootstrapping tests and a bit weaker in cross validation. However, as I pointed out earlier, accuracy scores are generally considered a bad measurement of model strength due to their sensitivity to the cutoff point and because they don't take

Table 24: Model C compared to models that exclude a single factor

Model	*Deviance compared to Model C*	*Bootstrapped C*	*Bootstrap Accuracy (n=200)*
Model C	306.28, df=10	0.935	89.86%
Model C - VClass	341.85, df=8, P<.001	0.922	88.19%
Model C - Length	333.21, df=9, P<.001	0.923	88.68%
Model C - NPType	318.23, df=9, P<.001	0.929	89.63%
Model C - Animacy	327.08, df=7, P<.001	0.927	89.01%
Model C - Access	340.49, df=6, P<.001	0.924	88.66%
Model C - Def	320.38, df=9, P<.001	0.926	89.1%

error/success margins into consideration. When we consider the corrected discrimination scores Model C is considerably stronger than Model B (these scores also remain about the same on a larger number of bootstrapping iterations).

Reestablishing the significant factors Upon arriving at a better model than our initial interactionless model, it is critical to reexamine which of the model's factors are significant. While the factors VClass, Length, NPType, Animacy, Access and Def were all found to be significant in Model A, the newly added interactions can theoretically invalidate some of them or perhaps strengthen their effect. In my the discussion of model A, the only metric used to conclude that a factor is significant was its contribution to the decrease in the model's deviance. Under this metric the factor Animacy and to an extent also Def were only marginally significant. In the discussion of Model C (below), more metrics are used, resulting in the conclusion that all factors are highly significant.

Model C contains six of our original factors and two interactions. To test the significance of the different factors, six new models were constructed, each containing all of Model C's factors but one. On removing a factor, all its interactions were also removed. For instance, removing the factor Animacy involved removing the interaction Animacy*Access (while at the same time keeping Access in the model of course). Each of the resulting models was examined with respect to 3 measurements: (i) the increase in deviance relative to Model C and its significance; (ii) the new C value corrected for overfitting (by 200 bootstrap iterations); and (iii) the accuracy averaged over 200 bootstrapping iterations. The results are summarized in table 24.

As can be seen in the table, removing any of the factors from Model C increases deviance significantly (P<.001). This result indicates that in the superior model C, all factors are highly significant (note that this is in opposition to Model A in which Animacy and to an extent Def, were only marginally significant). It can also be seen that removing any single factor from the model diminishes the model's

discrimination ability as is reflected by the lower C and accuracy scores (both corrected for overfitting by bootstrapping)[67]. The convergence of all measurements is encouraging and indicates the significant contribution of all factors to the strength of our final model.

5.3.3 Concluding Remarks

In this section I have examined the simultaneous influence of the different factors using several regression models and one classification tree model. The results obtained verify the second part of the topicality hypothesis (36-b) and outline a set of significant factors—VClass, Length, NPType, Animacy, Access and Def—the contribution of which to the choice of word order cannot be reduced to the effects of other factors.

Before concluding the discussion some caveats are in order. Firstly, the difficulty in directly quantifying topichood is problematic. Topichood is a factor in its own right and obviously one that should be included in the model[68]. Moreover, in the course of this work, I came across other factors that were not considered here and may very well change the results of this study, at least with respect to the exact set of significant factors. Verb tense, subject concreteness, structural parallelism, presence and location of specific modifiers and also a more fine grained classifications of the VClass and Access factors should all be considered in order to arrive at a more comprehensive picture. Furthermore, my modeling of interactions was rather basic; I did not consider third or higher order interactions, an analysis of which is likely to improve performance and provide more insight. Another important point to bear in mind is that the corpus used in this study was comprised entirely of written colloquial Hebrew. In order to generalize the obtained results to spoken Hebrew, the corpus has to be enriched with data of spoken Hebrew and the differences between the genres (differences that will most certainly arise) should be analyzed and accounted for.

Bearing all this in mind, the study reported here already provides a wealth of new data concerning the exact strength and influence of a large number of factors on the choice of word order. Indeed, it is at this point in time the only study that bears quantitatively on these issues. More importantly, I believe that this study is part of an important paradigm shift in the field of linguistics. Its results should not be taken to be the final word but rather a point of departure. In the emerging quantitative paradigm , if one wishes to contest my results by suggesting that other factors can better account for the data, all they have to do is to devise an empirical elicitation method for their factors, to construct a model that includes them and to provide the measurements indicating the superiority of their model.

[67] I conducted another 1000 bootstrapping iterations for Model C and its two closest "competitors". "Model C - NPType" and "Model C - Animacy". Results however, remained steady at C=0.935, 89.85%; C=0.93, 89.58%; and C=0.928, 88.99% respectively.

[68] Quantifying topicality is not impossible. Givón (1983) used a measurement that weighs the number of previous mentions (accessibility), the number of subsequent mentions (importance) and the number of distractors (other entities) in previous discourse (Givón's textual window is always 10 sentences before and after the sentence containing the entity in question). However, if we take topics to be aboutness topics it seems to me that Givón is just measuring topic correlates which is not that different than what I have been doing in this work (although I did not take into account all his topic correlates above, which is something that should be done by future research since they are indeed relevant). If one insists that topics should relate to our intuitive feeling of aboutness, an alternative way to identify the topic is to converge on a definition that yields high agreement rates between annotators and annotate sentences for the feature Top that will be 1 if the subject is topical and 0 otherwise. My conclusions from such an analysis based on Gundel's definition in (29) were that Top is a highly influential factor (arriving by itself to accuracy rates of 85.8% when fitted on the corpus data), but also that its inclusion in the model does not invalidate the other factors (although their degree of significance becomes lower, which is to be expected). I naturally don't take these data to be authoritative; conclusions can only be drawn after it is shown that Gundel's definition does indeed yield high agreement rates between different annotators. In the mean time the reader is referred to sections 3.3 and 3.4 for a qualitative discussion of the impact of topicality and why it does not tell the whole story.

Such methods bring linguistic methodology and argumentation to par with accepted standards in neighboring disciplines and are, to my view, an important and essential step forward.

6 Conclusions

In this work I have reviewed a number of current approaches to the analysis of Hebrew V1 sentences. I argued that the syntactic account presented by Shlonsky (1997) seems inadequate in the face of corpus data, and that the two major functional accounts for the phenomenon: Melnik (2002, 2006) and Kuzar (2006b, forthcoming), while mostly valid, appear to underestimate Givón's important generalization that all V1 constructions generally exhibit non topical subjects (Givón, 1976a). Based on this generalization and based on Lambrecht's evidence for its cross-linguistic relevance (Lambrecht, 1994, 2000), I have set forth an account that is based on two central claims: (i) that coding non-topical subjects is the driving force behind the availability of the V1 word order and its association with various linguistic features; and (ii) that since grammaticization is sensitive to frequencies, the various features associated with V1 constructions are at this point making an independent contribution to the choice of word order (beyond that of topicality). In order to substantiate my first claim I have examined the distribution of many linguistic features that are known to be topic correlates, validating that they all pattern as expected: features that are associated with low topicality are associated with the V1 word order and features that are associated with high topicality are associated with the S1 word order. In order to substantiate my second claim and to arrive at a set of significant factors I have constructed a number of statistical models to predict word order and showed that only a model that weighs the effects of multiple factors can account best for the corpus data. The model that provided the best results included factors relating to accessibility, verb class, animacy, definiteness, subject length and subject NP type. Attempts to remove any one factor from the model diminished its performance which indicates that all factors are required. I concluded the analysis in section 5.3.3 by suggesting empirical methods one can employ in order to falsify or (hopefully) add to these findings.

Part III

Appendixes

A The Syntactic Account of Triggered Inversion

Shlonsky's syntactic account of triggered inversion assumes that the clause-initial element in TI sentences is positioned in some specifier position within the CP, and carries with it a feature that must be checked. According to this account, the relevant feature must be checked by the verb, and for that reason the verb moves from I^0 to C^0 on top of its V^0 to I^0 movement. The V^0 to I^0 to C^0 movement places the verb hierarchically above the [spec, IP] subject, and yields the observed trigger-verb-subject word order.

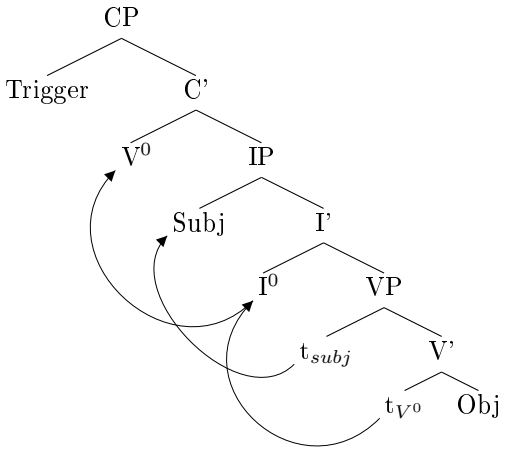

Figure 5: Simplified syntactic analysis of TI sentences according to Shlonsky (1997)

Shlonsky points out that practically any constituent that may appear clause-initially preceding the subject, can act as a trigger for inversion. Shlonsky (1997, p. 147) demonstrates many types of possible triggers, repeated below in (43):

(43) a. Temporal Adverb

etmol acra ha-mištara harbe pe'ilim.
yesterday detained the-police many activists.

'The police detained many activists yesterday.'

b. Prepositional Phrase

ba-pšita ha-leilit acra ha-mištara pe'ilim rabim.
in-the-raid the-nightly detained the-police activists many.

'The police detained many activicsts in the nightly raid.'

c. Clausal Adverb

mi-bli le-kabel išur mi-gavoha acra ha-mištara pe'ilim rabim.
without to-get authorization from-higher-up detained the-police activists many.

'The police detained many activists without getting authorization from higher up.'

d. Direct Object

pe'ilim rabim acra ha-mištara ba-pšita ha-leilit.
activists many detained the-police in-the-raid the-nightly.

'The police detained many activists in the nightly raid.'

e. Indirect Object

la-taxana šalxa ha-mištara et ha-acurim.
to-the-station sent the-police ACC the-detainees.

'The police sent the detainees to the station.'

f. Clausal Complement

lo le-daber be-mešex ha-nesi'a tav'a ha-mištara min ha-acurim.
NEG to-talk during the-ride demanded the-police from the-detainees.

'The police asked the detainees not to speak during the ride.'

g. Negative Phrase

le-olam lo taskim ha-memšala le-farek hitnaxaluyot.
never NEG will-agree the-government to-dismantle settlements.

'The government will never agree to dismantle settlements.'

h. Wh-Expression

matai acra ha-mištara et ha-pe'ilim?
when detained the-police ACC the-activists?

'When did the police detain the activists?'

Shlonsky does not go into details about the exact position of the different triggers inside the comp domain nor does he specify their relevant features. He abstracts from these details by saying that the triggers occupy different specifier positions within the CP (supposedly according to their role), and treats them all as [spec, CP].

Shlonsky then considers the VS word order in TI sentences to be motivated by the need to check features of the trigger. However, as discussed in Chapter 1, it is well known that inverted sentences normally have non-inverted alternatives. Treating inversion as an optional process poses two immediate problems for Shlonsky's ideas:

1. If inversion is optional, how can the relevant features of the trigger be checked in the case where inversion does not take place?
2. If these features do not require checking, then what motivates movement in the first place?

Shlonsky eschews these problems by arguing that inversion is in fact not optional but results from a blend of dialects:

> I believe the optionality in these cases reflects register or dialectal differences: formal written Hebrew requires inversion—this is particularly clear when the trigger is a wh-expression or a relative operator—while colloquial spoken Hebrew eschews it. In a strict sense, then, triggered inversion is not an optional process but results from a blend of dialects. (Shlonsky, 1997, p. 149)

However, a quick search in corpora of Spoken Hebrew reveals numerous examples of trigger-verb-subject sentences that are supposedly banned from Spoken Hebrew[69]

[69] The first example is from Dori-Hacohen's corpus of radio conversations (Dori-Hacohen, 2008) and the second and

(44) a. pit'om cacim kol miney anašim še-nora rocim miš'al am.
Suddenly appear all sorts of people who-terribly want a-referendum.
'Suddenly all sorts of people appear, who desperately want a referendum.'

b. pit'om ole l-o eyze fyuz l-a-roš.
suddenly rises to-him some fuse to-the-head.
'Suddenly he becomes very enraged.'

c. kax pit'om ba'a la eyze sirat mifras ktana ve-ata carix...
so suddenly arrive to-itself some sailboat small and-you need...
'So suddenly a small sailboat arrives and you need...'

It is therefore clear that inversion is not shunned by Spoken Hebrew, and at least synchronically, triggered inversion is an optional process. It might be argued that diachronically, the use of TI sentences in Spoken Hebrew originated from the influence of written texts, but this claim, beside being speculative, does not resolve the problem. The optionality of TI entails that synchronically, in the mind of the speaker, the features of the various triggers require checking on some occasions but not on others, which is of course contradictory.

third are from a subset of the Izre'el corpus Izre'el et al. (2004) obtained from the Mila knowledge center for processing Hebrew (http://www.mila.cs.technion.ac.il/hebrew/resources/corpora/spokenHebrew/index.html).

B Subject

B.1 Overview

In traditional grammar, the term 'subject' is highly ambiguous. Grammarians and logicians have used it to refer to quite distinct concepts such as "old information", "the thing which the sentence is about", "the prominent element in the sentence", etc. Jespersen (1924, p. 146), surveyed these definitions and concluded that it is best to restrict the word subject to refer to the grammatical subject (see below), and use different terms for the other concepts.

It is interesting to find the exact terms used in current discussions of topicality as the focus of attention in papers written some 100 years ago on subjecthood. For instance, Jespersen (1924) quotes Baldwin (1902, p. 364):

> The subject is sometimes said to be the relatively familiar element, to which the predicate is added as something new. The utterer throws into his subject all that he knows the receiver is already willing to grant him, and to this he adds in the predicate what constitutes the new information to be conveyed by the sentence.

Later he writes that another "frequently given" definition is that the subject is what you talk about, and the predicate is what is said about the subject. These two definitions, using the concepts of aboutness and givenness, have survived to this day, but following Jespersen's recommendations they are no longer used to define subject but rather to describe the characterizations of the topic (see appendix C).

Jespersen goes on to survey the terms *logical subject* and *psychological subject*. These terms were put forth in an attempt to clarify different aspects of subjecthood, but have soon become hopelessly ambiguous themselves. The Oxford Dictionary of English Grammar (Chalker and Edmund, 1998) currently defines the psychological subject as "what the clause is about", and the logical subject as "the agent of the action". However, in the increasingly rare occasions where these terms are used, their definitions vary considerably. Jespersen (1924) listed no less than eleven definitions for these terms, where only two of which are similar to those of the Oxford Dictionary[70].

Jespersen concluded his survey in saying that we should restrict the use of the term subject to refer to the grammatical subject, and avoid attaching to this word adjuncts such as logical or psychological. It is interesting to note that Jespersen's own definition for grammatical subject in Jespersen (1924) has not gained popularity over the years[71]. However, in Jespersen (1937, p. 137) he characterized it by verbal agreement which is similar to the way in which it is perceived today (see discussion in the following section).

[70] The definition of psychological subject by Paul (Jespersen, 1924, p. 147), reminds in some ways the modern discussion of topics. Paul considered the psychological subject to be the idea or group of ideas that is first present in the mind of the speaker and the psychological predicate to be the part that is later joined to it. He also said that the speaker sometimes places the psychological subject after the predicate, because in the moment where he begins to speak, the predicate idea is the more "important" one. While this wording is far from the one used today, it might be considered an early discussion of the factors affecting information packaging. As for the Oxford Dictionary's definition of logical subject, it can also be tracked back to Jespersen (1924), where along with other definitions it was claimed that "Many grammarians use the term *logical subject* for that part of the passive sentence which would be the subject if the same idea had been expressed in the active turn".

[71] Jespersen (1924) attempted to outline a number of grammatical manifestations that characterize the "primary" element in the sentence resulting in a mixture of semantic concepts that is quite distinct from what is termed grammatical subject in modern linguistics.

B.2 Grammatical Subject

The grammatical subject is so named because it is grammatically marked in a way that differentiates it from other elements of the sentence. Theories differ with regard to the functional role of subject marking. Mithun (1991, p. 160), for example, argues that "the function of subjects is clear: they are essentially grammaticized clause topics." Dowty (1991) on the other hand presented a granular semantic theory in which the subject role is to mark the participant that is the most "agentive", that is, the one with more agentive characteristics like volition or causality[72]. It seems reasonable to assume that topichood and agentivity (in the sense of the proto-agent roles) are both different aspects of subjecthood. If we look for instance on passive sentences such as *John was shot by a sniper*, it seems that the choice of John for the subject is due to his topic status (he has less proto-agent roles than the sniper). However, if we look at a sentence such as *koev l-i ha-roš* 'hurt to-me the-head' it seems that in this Hebrew sentence the subject *ha-roš* 'the-head' is selected not because of its topichood (it is far more likely that the speaker is the topic in this sentence) but rather because of its proto-agent roles (in this case causality). These data demonstrate an important issue discussed by Evans and Levinson (in press). In the process of formulating sentences, many different aspects of the underlying propositions compete for coding. Coding too many of these will yield cumbersome sentences, but coding only a few will make our sentences less effective and harder to process. In this respect the subject can then be seen as a clever grammatical mechanism to uniformly code propositional aspects that correlate statistically. Instead of separately coding topichood and each of Dowty's roles, our grammar notes that they often coincide and codes the element that embodies them best as the grammatical subject. As to the question of which element is "the element that embodies them best", I tend to think a simple counting mechanism like the one suggested in Dowty (1991) is over simplistic (see for instance the passive example above). The effect of the different aspects on the choice of subject should in itself be analyzed using multifactorial techniques, but such investigation is outside the scope of this work. Following the above discussion, I will use the term *primary participant* as a cover term for "the element of the proposition in which aspects such as topichood and proto-agenthood interact more strongly than in other sentence elements[73]", and I will consider the role of the grammatical subject to be the coding of this element.

As for the grammatical manifestation of subjects, sidestepping some cross-linguistic variation subjects are usually marked by at least one of the following three features: (i) word order, (ii) agreement with the predicate, and (iii) nominative case[74]. In Hebrew, subjects are prototypically marked by all three, but the picture is not always that simple. Inverted sentences, which are the topic of this work, are not uncommon, so in Hebrew, word order is the least reliable subject marker. To complicate things further, we often encounter inverted [V S] sentences in which it is not only that the subject does not precede the verb, but also agreement is lost, or worse yet, the accusative case is assigned instead of the nominative. This difficulty is best exemplified by sentences containing the possession/existence predicate *yeš:*

[72] Dowty (1991) presented a list of 5 proto-agent and 5 proto-patient roles. His proto-agent roles are: volition, sentience, causation, movement and independent existence and his proto-patient roles are change of state, incremental theme, causally affected, stationary relative to another participant and existence not independent from the event. He argued that the subject's role is to mark the participant with the greatest number of proto-agent roles in the event.

[73] The exact definition of "strongly" is language specific and should be determined by multifactorial analysis. The exact set of relevant aspects of the proposition as well as their exact degree of influence on determining the primary participant is outside the scope of this work.

[74] Nominative case is partially expressed in Hebrew by the lack of the accusative case on definite NPs (i.e. the absence of the accusative marker *et*).

(45) a. yeš et ha-sfarim(3PM) b-a-sifriya.
EXIST ACC the-books(3PM) in-the-library.
'The books are in the library.'

b. yeš l-i et ha-sfarim(3PM) b-a-dira.
EXIST to-me ACC the-books(3PM) in-the-apartment.
'I have the books in my apartment.'

c. yeš šomeret(3SF) b-a-knisa.
EXIST a-guard(3SF) in-the-entrance.
'There is a guard at the entrance.'

What is the grammatical subject of the sentences in (45)? The only argument of the predicate in the sentences above is post verbal, it does not agree with the verb and it is marked by the accusative case.

Ziv (1976) examined the grammatical markings of subjects in Hebrew possessive sentences (similar to sentence (45-b)) and concluded that Spoken Hebrew subjects of this type have already lost the subject position and case marking features of canonical subjects and are in the process of losing agreement as well[75]. She argues that these findings are in accord with Keenan's promotional hierarchy (Keenan, 1976) according to which from the three topic coding properties position is easiest to lose, followed by case marking and followed by verbal agreement. She concludes by arguing that in Spoken Hebrew possessive sentences both the possessor and the possessed element can assume different subject properties[76], and there is in fact a process of reanalysis of the grammatical relation between these two elements. Nonetheless, in practically all other cases beside those involving the predicate *yeš* , Hebrew subjects maintain at least one of the subject properties (i.e. either case or agreement, see also section 5.2.1)[77] . I will therefore define the subject in Hebrew as the element of the sentence that either agrees with the verb or is assigned the nominative case (or both of course). I will concede however, that in the specific case of the existence predicate *yeš,* this definition is insufficient and because of the lack of a clearly superior alternative I will take the subject of this predicate to be the element whose existence/possession is being asserted.

[75] Ziv examined sentences with the past and future inflections of the existence predicate *yeš* 'there-is' (i.e. *haya* 'there-was' and *yihiye* 'there-will-be'). Spoken Hebrew sentences with these verbs have accusative marked, post verbal subjects, although subject–verb agreement is still sporadically maintained. Ziv compared this situation to that of Normative Literary Hebrew where possessive subjects are also post verbal, but they still agree with their predicate and maintain the nominative case. From this comparison she concluded that diachronically, Hebrew is in the process of losing the subject markers in possessive sentences.

[76] The possessor can appear sentence initially and thus assume the word order characteristic of subjects. It however never assumes the other two characteristics—verbal agreement and nominative case.

[77] There are a handful of other predicates that might also, very rarely, appear without a subject-marked argument. These predicates are semantically very similar to the existence predicate (e.g. *kara(3SM) et hamikrim(3PM)* 'happened(3SM) ACC the-incidents(3PM)'), but these cases are limited to informal registers, and are rare even there. My sample of the Linzen corpus (370 V1 sentences and 561 sentences overall) did not include any such cases.

C Sentence Topic

C.1 Overview

The term *sentence topic* is commonly defined as the subject-matter or "what the sentence is about" (Hocket 1958:21). The precise characterization of this concept is subject to great controversy, but it is nevertheless repeatedly alluded to when discussing issues of word order, or broadly speaking, when discussing packaging variants for propositions. Attempts to formalize aboutness led to definitions in terms of: givenness, limiting the predication domain, mental addressation, information gain, and also, many researchers just stick to the vague concept of aboutness in lack of a clearly superior one. In this appendix I will introduce the motivation behind this concept, discuss its phenomenology and a number of approaches to its definition. I will conclude by adopting a definition of topic that is based on Gundel (1988) and Reinhart (1981).

C.2 Topic Phenomenology

The term topic phenomenology, refers to the range of phenomena that has been motivated, at least to some extent, by reference to the concept of topic. These phenomena include a plethora of syntactic constructions that are argued to be motivated by topic coding, as well as some other phenomena mentioned below. A syntactic structure is said to be motivated by topicality insofar as it encodes pragmatic structure that relates to topicality (or the lack thereof). Let us then briefly discuss the possible pragmatic structures of propositions and their accompanying sentence structures.

The unmarked structure of propositions is *topic-comment*. That is, the *topic* is the element that the proposition is about, and the *comment* is the assertion made about that topic. Propositions can also be structured as a single unit without topic-comment relations—that would be the case in event reporting sentences such as *it's raining*. Sentences reflect the internal pragmatic structure of their underlying propositions. That is, sentences that encode topic-comment relations will have a different form than those encoding topicless propositions. Furthermore, sentences coding topic-comment relations also vary in form from one another, depending on whether the subject coincides with the topic, the degree of topic activation, etc.

As argued in appendix B.2, the grammatical subject codes the role of the primary participant, which most of the time (but not always), coincides with the topic. As a result, a language's canonical word order, if it has one, is usually also its main mechanism for topic coding. The sentences in (46) are some Hebrew examples of prototypical and non-prototypical topic coding constructions.

(46) a. Prototypical Topic Coding

Dan pirsem moda'at drušim b-a-iton.
Dan published an-add wanted in-the-newspaper.

'Dan published a want ad in the newspaper.'

b. Passive Construction

modaot drušim hitparsemu b-a-iton.
ads wanted were-published in-the-newspaper.

'Want ads were published in the newspaper.'

c. Topicalization

et moda'at ha-drušim pirsem dan.
ACC ad the-wanted published dan.

'The want ad was published by Dan.'

d. Left dislocation

ba-ašer le-moda'at ha-drušim, pirsem ota dan.
with-regard to-the-ad the-wanted, published it dan.

'With regard to the want ad, Dan published it.'

e. Hanging topic

ba-ašer le-peša, ani maskim im sar-hapnim.
with-regard to-crime, I agree with the-minister-of-interior.

'With regard to crime, I agree with the Minister of Interior.'

Sentence (46-a) is a canonical subject-predicate/topic-comment sentence. (46-a) is unmarked in the sense that it is in the active voice, and its subject is also the topic. Sentences (46-b) (46-c) and (46-d) all exhibit well known strategies of marking topical objects, and sentence (46-e) is distinct from the others in the sense that its topic *crime* is separated from the clause encoding the comment.

Beside the canonical topic constructions above, Hebrew also uses verb first (V-Object-S) sentences to code topical objects. This structure can arise with direct, indirect and dative or locative objects, as exemplified below (the topical objects are in bold):

(47) a. Direct Object

acar oti šoter.
arrested me a-policeman.

'A policeman arrested me.'

b. Indirect Object

hitxil iti mišehu b-a-mesiba.
flirt with-me someone at-the-party.

'Someone at the party made a pass at me.'

c. Dative Modifer

koev li ha-roš.
hurt to-me the-head.

'I have a headache.'

d. Locative Modifier

nafla po pcaca.
fell here a bomb.

'A bomb fell here.'

Finally, topicless propositions are often coded in VS sentences. Depending on discourse context, such sentences might contain objects (see discussion of thetic propositions it section 2.2.2), but the prototypical examples have only a subject and a predicate.

(48) Thetic propositions

a. yored gešem.
falling rain.

'It's raining.'

b. yeš makot.
EXIST a-fight.
'There's a fight.'

It should be noted, that topicless propositions and propositions with topical objects can also be encoded in the unmarked SV(O) word order and indeed they often are[78]. The inverse is very rare, that is, the marked constructions in (46), (47) and (48) will not normally encode canonical topic-comment propositions in which the topic is also the subject.

Beside the above constructions, many other linguistic phenomena were explained with reference to topicality. These include phenomena that have been widely discussed in modern linguistics such as anaphora resolution, the dative alternation and even island constraints (cf. Reinhart, 1981, 1983, Erteschik-Shir, 2007). A lot of the research in the field relies on some intuitive sense of aboutness to motivate some specific construction or phenomenon. Attempts to explicate these intuitions in a way that would accommodate the wide range of phenomena have encountered considerable difficulty. In fact, some recent research questions the validity of the concept altogether (cf. Bar-Asher, 2009, Jacobs, 2001). In the following section I will review different proposals and devise a working definition of topic.

C.3 Aboutness and Givenness in the Definition of Topics

Attempts to define the term topic have often involved the concepts of aboutness and givenness. Indeed, it can be said that most current approaches to topicality can be classified by their attitude toward these two concepts. Most researchers agree that aboutness should be a defining feature of topics, but they often differ on their concept of aboutness: some take it as a primitive while other explicate it further. As for givenness, opinions vary even more. Some argue that the topic should be active in the mind of the hearer prior to the utterance (Strawson, 1964), some say it is sufficient for it to be familiar (Gundel, 1988, Gundel and Fretheim, 2001), while others argue that topics can be non-familiar (Reinhart, 1981, Michaelis and Francis, 2007).

A common way to introduce topics in terms of givenness is the following[79]:

(49) The topic is the part of the proposition that is given, i.e. it is known to both the speaker and the hearer. The comment is the new information added about the topic.

As argued by Gundel (1988), such characterization subtly conflates two senses of givenness-newness: relational and referential.

Referential givenness-newness involves a relation between a *linguistic expression* and a corresponding *non-linguistic entity* in the speaker's/hearer's mind. The status of referential givenness is the degree of activation the non-linguistic entity has in the mind of the hearer at the onset time of the utterance. It can be, for instance, active (just mentioned in discourse or otherwise salient from the speech settings), familiar, identifiable or it can be brand-new and unidentifiable.

[78] That's true for most types of propositions. although the word order of canonical thetic propositions such as the ones in (48) is already fixed on VS.

[79] This formulation is essentially the one quoted from Baldwin (1902) in section B, but it is very common and by no means limited to that text.

Relational givenness-newness, involves the partition of the proposition into two complementary parts, X and Y, where X is what the sentence is about, and Y is what is predicated about X. Y is new **in relation** to X in the sense that it adds new information about it; X is given **in relation** to Y for the same reason. Relational givenness-newness is in fact a way to characterize aboutness through information gain, and it is a different concept from the one commonly referred to by givenness.

Coming back to the definition in (49), we can now see the fallacy. The definition supposedly creates a givenness-newness contrast between the topic and the comment, while in fact, these are different senses of givenness-newness. When discussing the topic, the definition in (49) alludes to referential givenness (knowledge in the mind of the hearer), while when discussing the comment it alludes to relational newness (the information is only new in relation to the topic, it can, and often does, contain referentially old entities). Such a definition conflates aboutness and givenness in a way that suggests that if the comment is new information **in relation** to the topic, the topic ought to be given information **in the speaker's mind**. In actuality, the concepts are distinct. Take for instance the examples in (50) where the subject-topics are underlined:

(50) a. ... The public benches that used to be west of their restaurant are gone also. It has been rumored that the removal of the benches has been brought about by pressure from certain business people who want to discourage those who can't afford to get drunk in public behind iron work railings, from annoying those who can. Of course, one of consequences is that the tenants of 1415 Ocean Front Walk don't have their benches to sit on...[80]

b. She sent him to kindergarden. As soon as he went there, the teacher took one look at him and he threw up again.[81]

c. etmol b-a-boker ra'iti ka'amur et "adama mešuga'at. [...].
yesterday in-the-morning I-saw as-previously-said ACC "earth crazy". [...].
ronit yudkevič mesaxeket madhim b-a-seret ha-ze ve-zo l-i
Ronit Yudkevich acts ammazingly in-the-film that and-it-is to-me
ha-pa'am ha-rišona še ani ro'e ota be-seret yisraeli. saxkanit le-eyla u-le-eyla.
the first time that I see her in-a-movie Israeli. Actress wonderful.
'Yesterday morning I saw the movie "Sweet Mud". Ronit Yudkevich is amazing in that film. What a wonderful actress.'

Reinhart (1981) and Michaelis and Francis (2007) argued based on sentences (50-a) and (50-b) (among others) that the element that the sentence is about (hereafter *the aboutness topic*) can be discourse new[82]. Another example is (50-c) from the Linzen's Hebrew blogs corpus (Linzen, 2009). In all these cases it can be argued that the aboutness topic is primed by previous discourse but there's no doubt it is not referentially given.

While I generally agree with Reinhart and Michaelis & Francis that a high degree of givenness is not **required** of topics, I should point out that givenness undoubtedly contributes to our intuitive feeling that the sentence is about a certain entity; if a sentence has both a referentially new entity and a referentially old one, we would be **more likely**—all other things being equal—to judge the referentially old one as topical[83]. In this sense, I believe that givenness influences aboutness, but that

[80] Reinhart (1982:21) from a magazine article

[81] Michaelis & Francis (2007:24), from the switchboard corpus of English telephone conversations.

[82] I do not have enough sentence context in order to determine if *the teacher* in (50-b) is indeed topical, but that was the judgment of Michaelis & Francis. I bring this sentence mostly to credit the authors; examples of discourse new topical subjects are not hard to come by.

[83] The sentences in (50) all show that this is not always true. In all sentences the underlined topical element appeared

there is no hard constraint on the givenness of the aboutness topic.

Bearing in mind that aboutness and givenness are two related but distinct concepts, it is still possible to combine the two in the definition of topic. Strawson's definition (Strawson, 1964, p. 97,98) explicitly stated both conditions:(i) the topic is what the statement is about, and (ii) the topic is used to invoke "knowledge in the possession of an audience.[84]"

Lambrecht's definition (Lambrecht, 1994, p. 131) is harder to pin down:

> A referent is interpreted as the topic of a proposition if in a given situation the proposition is construed as being about this referent, i.e. as expressing information which is relevant to and which increases the addressee's knowledge of this referent.

Lambrecht supposedly defines topic through aboutness alone, but in fact, his notion of relevance seems to entail a high degree of givenness. In (Lambrecht, 1994, p. 119) he appears to follow Strawson's "Principle of Relevance" stating that the topic is "a matter of standing interest or concern".

Such a dual approach to the definition of topic has at least one practical advantage—it allows for an easier identification of the topic. Aboutness is a vague concept, and it is often hard to identify the topic relying on aboutness alone. Many times, the difficulty in deciding the topichood status of entities comes hand in hand with the low givenness of these entities. In these cases, a requirement for a high degree of givenness will immediately render the element non-topical. In practice, such an approach will assist in getting higher agreement rates from different informants on the task of topic identification.

If aboutness+givenness were to account for the range of phenomena commonly associated with topic, then on the basis of the practical consideration alone it might have been desirable to define the topic that way. However, it appears that such a definition runs into problems with some of the prototypical topic constructions discussed in (46). Notably, left dislocation, which normally involves dislocating an element that is familiar and discourse old but not active, would have to be considered a focus device rather than a topic device. This result seems undesirable. Left dislocation constructions are probably the prime examples of aboutness (e.g. they can open with "and about X ..."), and if we exclude them from our phenomenology then we should also exclude hanging topics and we appear to lose too much.

On this basis, and from considerations of simplicity and elegance (theoretical machinery should not be added where it is not empirically needed), I believe the correct approach is to define the topic in terms of aboutness alone and not constrain it by givenness. The givenness restrictions on topichood should not be a primitive but rather an empirical question. This approach is espoused by both (Gundel, 1988, Gundel et al., 1993, Gundel and Fretheim, 2001) who argues for a familiarity constraint on aboutness topics, and Reinhart (1981) that sets the bar even lower, at non-familiar specific indefinites, excluding only non-specific indefinites (see her example in (50-a) above. (50-c) also exemplifies this point).

Following Reinhart (1981) and Gundel (1988) I will assume that the topic is "what the sentence is about" and I will equate topic with aboutness topic. To explicate aboutness I will be using Gundel's topic definition (Gundel, 1988, p. 210) that is designed to formulate the intuitive sense of the concept.

along with another non-topical element that had higher referential givenness. While referential givenness is a strong influence on our judgments of aboutness, there are other factors and obviously the grammatical SVO structure and the high degree of animacy of the entities involved played a critical role in the sentences in (50). For further discussion see section 3.5.

[84] the audience in this case being the hearer. Strawson second condition is basically givenness. Strawson also required a high level of givenness, as he argued that the topic must be "a matter of standing interest or concern."

(51) Topic Definition: An entity, E, is the topic of sentence, S, iff in using S the speaker intends to increase the addressee's knowledge about, request information about, or otherwise get the addressee to act with respect to E.

This definition, while still intuitive, is an important step forward. Another helpful way to identify topics, is through Reinhart's catalog metaphor. Reinhart (1981) compares the speaker and the hearer's representation of the discourse context to a list of propositions they consider true—their context set. Reinhart goes on to argue that in much the same way that library books are indexed by author or title, the propositions in our discourse context are indexed by topic. Once the hearer encounters a new sentence, he identifies its topic and "catalogs" the proposition under its entry in the context set. If the proposition is topicless, it remains uncatalogued (supposedly in a list of topicless propositions). Within this metaphor, the topic is seen as an instruction from the speaker to the hearer to catalog a proposition under a specific context set entry[85].

The most prevalent argument against the metaphorical explications of aboutness regards their vagueness. Maslova and Bernini (2006, p. 5) dismiss these approaches with a footnote saying "The explication of aboutness in terms of mental addressation seems to be based on overly simplistic and metaphorical model of human memory and, in fact, does not provide any more explicit criteria for identification of topics than the intuitive notion of aboutness." they then go on to take aboutness as a primitive in their characterization of topic.

Personally, I tend to disagree with Maslova & Bernini's above statement. I have tried working with non-linguist informants presenting them with sentences and asking their judgments on the task of identifying the element the sentence is about. When I appealed to their natural understanding of "aboutness", their judgments were more scattered and less similar to those of a linguist than when presented with an explication of the term by Gundel's definition and Reinhart's metaphor. Nevertheless, I do concede that these explications of aboutness are still too vague to support stringent empirical research. For this reason, in the empirical part of this work I have used topic correlates to bear out the role of topicality (see section 3.5 and part II). However, in order to fully validate the conclusions of this work—and indeed the conclusions reached in other studies which assume the existence of the topic category—an agreed upon empirical elicitation method for topics is needed. Further avenue for research would be to take an intuitive definition such as Gundel's, and to empirically show that it can yield high agreement rates among non-linguists when judging various topic related phenomena (see Dabrowska (2009) on the importance of working with non-linguist informants).

[85] I do have one reservation about Reinhart's metaphor and it relates to her view that a sentence can have at most one sentence topic. At least where V1 constructions are involved, I do not see a need to impose this restriction. The question relevant for V1 constructions is whether the subject is topical, the presence of other topical elements does not seem to affect word order. Furthermore, there is evidence from other constructions that allowing multiple topics can be beneficial (the English dative shift is one such example, see Givón (1979, p. 160,161)). While I still use Reinhart's metaphor when identifying topics, I consider it possible to catalog a single sentence under more than one topic.

Bibliography

Alexiadou, A., 2007. Post-verbal nominatives: an unaccusativity diagnostic under scrutiny. In Conference on Linguistic Interfaces, University of Ulster.

Ariel, M., 1988. Referring and accessibility. *Journal of linguistics*, volume 24(1):65–87.

Ariel, M., 1990. Accessing noun-phrase antecedents. Routledge (London, New York).

Ariel, M., 2001. Accessibility theory: an overview. *Text representation: Linguistic and psycholinguistic aspects*, pages 29–87.

Arnold, J., Wasow, T., Losongco, A., et al., 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, volume 76(1):28–55.

Baayen, R., 2008. Analyzing linguistic data: A practical introduction to statistics using R. Cambridge Univ Pr.

Baldwin, J., 1902. Dictionary of Philosophy and Psychology, volume 2. The Macmillan company.

Bar-Asher, E., 2009. A Theory of Argument Realization and its Application to Features of the Semitic Languages. PhD diss., Harvard University.

Belletti, A., 1988. The case of unaccusatives. *Linguistic inquiry*, volume 19(1):1–34.

Borer, H., 2005. The Normal Course of Events (Structuring Sense, vol. II).

Brentano, F., 1874. Psychology from an empirical point of view. *Translated by Antos C. Rancurello, DB Terrell, and Linda L. McAlister from Psychologie vom empirischen Standpunkt*, volume 1924.

Bresnan, J., Cueni, A., Nikitina, T., et al., 2007. Predicting the dative alternation. *Cognitive foundations of interpretation*, pages 69–94.

Chalker, S. and Edmund, W., 1998. The Oxford Dictionary of English Grammar. Oxford University Press.

Chomsky, N., 1981. Lectures on Government and Binding. Foris, Dordrecht.

Comrie, B., 1981. Language universals and language typology. *Chicago: University of*.

Dabrowska, E., 2009. Naive v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review*.

Deane, P., 1992. Grammar in mind and brain: explorations in cognitive syntax. Walter de Gruyter.

Dori-Hacohen, G., 2008. Corpus of radio conversations. Collected for a Ph.D. dissertation at Haifa University supervised by Katriel, T. and Maschler, Y.

Dowty, D., 1991. Thematic proto-roles and argument selection. *Language*, volume 67(3):547–619.

Erteschik-Shir, N., 2007. Information structure: the syntax-discourse interface. Oxford University Press, USA.

Evans, N. and Levinson, S., in press. The myth of language universals: Language diversity and its importance for cognitive science. *Brain and Behavioural Sciences*.

Giora, R., 1981. Sentence ordering as a text dependent phenomenon. In S. Blum Kolka, Y. Tobin, and R. Nir, editors, Studies in Discourse Analysis, pages 264–302. Jerusalem: Academon.

Givón, T., 1976a. On the VS order in Israeli Hebrew. *Studies in Modern Hebrew Syntax and Semantics*, (32).

Givón, T., 1976b. On the VS word order in Israeli Hebrew: pragmatics and typological change. *Studies in Modern Hebrew Syntax and Semantics: The Transformational-Generative Approach*, page 153.

Givón, T., 1976c. Topic, pronoun and grammatical agreement. *Subject and topic*, pages 149–188.

Givón, T., 1979. On understanding grammar. Academic Pr.

Givón, T., 1983. Topic continuity in discourse: A quantitative cross-language study. J. Benjamins.

Givón, T., 1992. On interpreting text-distributional correlations: Some methodological issues. *Pragmatics of word order flexibility*, pages 305–320.

Goldberg, A., 1995. Constructions: A construction grammar approach to argument structure. University of Chicago Press.

Goldberg, A., 2006. Constructions at work: The nature of generalization in language. Oxford University Press, USA.

Gries, S., 2003. Multifactorial analysis in corpus linguistics: A study of particle placement. Continuum Intl Pub Group.

Gundel, J. and Fretheim, T., 2001. Topic and focus. *Handbook of Pragmatic Theory. Oxford: Blackwell.*

Gundel, J., 1988. Universals of topic-comment structure. In M. Hammond, E. Moravcsik, and J. Wirth, editors, Studies in syntactic typology, pages 209–239. Amsterdam: John Benjamins.

Gundel, J., Hedberg, N., and Zacharski, R., 1993. Cognitive status and the form of referring expressions in discourse. *Language*, volume 69(2):274–307.

Harrell, F., 2001. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Springer Verlag.

Hawkinson, A. and Hyman, L., 1974. Hierarchies of natural topic in Shona. *Studies in African Linguistics*, volume 5(2):147–169.

Izre'el, S., Hary, B., Du Bois, J., et al., 2004. The corpus of Spoken Israeli Hebrew. URL `http://www.mila.cs.technion.ac.il/hebrew/resources/corpora/spokenHebrew/index.html`.

Jacobs, J., 2001. The dimensions of topiccomment. *Linguistics*, volume 39(4):641–681.

Jespersen, O., 1924. The philosophy of grammar. London: Allen & Unwin.

Jespersen, O., 1937. Analytic Syntax. London: Allen & Unwin.

Keenan, E., 1976. Towards a universal definition of subject. *Subject and topic*, pages 303–333.

Kuroda, S., 1972. The categorical and the thetic judgment: Evidence from Japanese syntax. *Foundations of language*, volume 9(2):153–185.

Kuzar, R., 1990. Message Structure of the Sentence in Israeli Hebrew [in Hebrew]. Ph.D. Dissertation.

Kuzar, R., 2006a. Sentence Patterns in Israeli Hebrew according to Rosen. *Hebrew and its Sisters*, pages 269–294.

Kuzar, R., 2006b. The Existential Construction as Contributor to the Existential Meaning. *Language Studies*, pages 101–112.

Kuzar, R., forthcoming. Sentence Patterns in English and Hebrew.

Lakoff, G., 1987. Women, fire, and dangerous things: What categories reveal about the mind. University of Chicago press Chicago.

Lambrecht, K., 1994. Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents. Cambridge University Press.

Lambrecht, K., 2000. When subjects behave like objects: An analysis of the merging of S and O in sentence-focus constructions across languages. *Studies in Language*, volume 24(3):611–682.

Levin, B. and Rappaport Hovav, M., 1995. Unaccusativity: at the syntax-lexical semantics interface. MIT Press.

Linzen, T., 2009. Corpus of blog postings collected from the Israblog website.

Marty, A., 1918. Gesammelte Schriften, vol. II. *Halle, Max Niemeyer.*

Maslova, E. and Bernini, G., 2006. Sentence topics in the languages of Europe and beyond. *Pragmatic organization of discourse in the languages of Europe*, volume 20:8.

Melnik, N., 2002. Verb-Initial Constructions in Modern Hebrew. Ph.D. thesis, University of California at Berkeley.

Melnik, N., 2006. A constructional approach to verb-initial constructions in Modern Hebrew. *Cognitive Linguistics*, volume 17(2).

Michaelis, L. and Francis, H., 2007. Lexical subjects and the conflation strategy. *The Grammar-Pragmatics Interface: Essays in Honor of Jeanette K. Gundel*, page 19.

Mithun, M., 1991. The role of motivation in the emergence of grammatical categories: the grammaticization of subjects. *Approaches to grammaticalization*, volume 2:159–184.

Preminger, O., 2009. Failure to agree is not a failure: phi-agreement with post-verbal subject in Hebrew.

Prince, E., 1981. Toward a taxonomy of given-new information. *Radical pragmatics*, pages 223–255.

Prince, E., 1992. The ZPG letter: Subjects, definiteness, and information-status. *Discourse description: Diverse linguistic analyses of a fund-raising text.*

R Development Core Team, 2009. R: A language and environment for statistical computing. *Foundation for Statistical Computing, Vienna, Austria.*

Reinhart, T., 1981. Pragmatics and linguistics. *An analysis of sentence topics. Philosophica*, volume 27:53–94.

Reinhart, T., 1983. Anaphora and semantic interpretation. Taylor & Francis.

Reinhart, T. and Siloni, T., 2004a. Against the unaccusative analysis of reflexives. *The unaccusativity puzzle: Explorations of the syntax-lexicon interface*, pages 159–180.

Reinhart, T. and Siloni, T., 2004b. Against the Unaccusative Analysis of Reflexives. *The Unaccusativity Puzzle*, pages 288–331.

Sasse, H., 2006. Theticity. *Pragmatic Organization of Discourse in the Languages of Europe*, page 255.

Shlonsky, U., 1987. Null and displaced subjects.

Shlonsky, U., 1997. Clause Structure and Word Order in Hebrew and Arabic an Essay in Comparative Semitic Syntax. Oxford University Press.

Sornicola, R., 2006. Interaction of syntactic and pragmatic factors on basic word order in the languages of Europe. *Pragmatic Organization of Discourse in the Languages of Europe*, page 357.

Strawson, P., 1964. Identifying reference and truth values. *Theoria*, volume 30(2):96–118.

Taub-Tabib, H., 2007. The Strong vs. Weak Unaccusative Hypothesis: Experimental Investigation. *Unpublished manuscript.*

Timberlake, A., 1975. Hierarchies in the genitive of negation. *Slavic and East European Journal*, pages 123–138.

Uhlířová, L., 1969. Vztah syntaktické funkce větného členu a jeho místa ve větě (Metody a výsledky statistického zkoumání). *Relationship between syntactic function and linear position of sentence elements (Methods and results of statistical research)/, Slovo a slovesnost*, volume 30:358–370.

Williams, R., 1994. A Statistical Analysis of English Double Object Alternation. *Issues in Applied Linguistics*, volume 5(1):37–58.

Ziv, Y., 1976. On the reanalysis of grammatical terms in Hebrew possessive constructions. *Studies in modern Hebrew syntax and semantics: The transformational-generative approach*, pages 129–152.