Stochastic Processes in Language: A Classical Quantification of the Anglo-Estonian Dialect

Kalev Hannes Leetaru

Friday, March 9th, 2001

INTRODUCTION AND BACKGROUND

Stochastic processes form the fundamental basis of much of modern information theory. In any human-derived language, a distinct linear selection process is undertaken in the construction of the arbitrary sentence. Depending on the thought to be expressed, a definite preference is granted to the particular word used to instantiate the sentence. This preference, in turn, tends to discriminate the available choices for the subsequent word, which does the same for the third word, ad infinitum, for the duration of the sentence. Such a stochastic process makes quantitative analysis a straightforward affair.

Many sociolinguistic factors are influential in this process, and an analysis of such factors is beyond the scope of this paper. Instead, a classical quantification of the Anglo-Estonian dialect is presented, with features and methodologies drawn from the domain of stochastic information theory. As such, the conclusions obtained are not intended to be encompassing of the domain, but rather objectively observative of the selected dialect.

METHODOLOGY

The fieldwork for the analysis portions of this research was conducted during the week of February 25th, 2001 and consisted of the analysis of the speech patterns of a single male subject. The selection criteria used was a demonstrated fluency in the Ango-Estonian dialect. The subject selected was a forty-seven year-old

male born in the New York City region of the United States of America and in possession of American citizenship. Raised by Estonian parents who spoke only Estonian at home, the subject was encouraged in the use of both languages, but learned Estonian before English.

The subject was engaged in a free-formed discussion of his favorite topic and was recorded for six minutes and seventeen point thirty-six seconds onto analog audio media. After recording, the speech was orthographically and phonetically transcribed and analyzed to provide quantification factors. A copy of the complete orthographic transcription is provided as Appendix I in this paper.

MORPHOLOGICAL ANALYSIS

The first stage of linguistic analysis performed was a morphological one. A selection of ten polymorphemic words was taken from the transcription corpus and analyzed. A summary of the analyses appears in Table I below. An "(I)" beside a delineation structure represents an inflectional affix, while a "(D)" beside a delineation structure represents a derivational affix.

TABLE I

Word	Morphological Constituent Structure	
	Graph Form	**Delineation**
Demeaning	[[de [mean]] ing]	Root: mean = Verb De + mean => Verb (D) Verb + ing => Adjective (I)
Independent	[in [[de [pend]] ent]]	Root: pend = Verb De + pend => Verb (D) Verb + ent => Adjective (D) in + Adjective => Adjective (D)
Surviving	[[survive] ing]	Root: survive = Verb Survive + ing => Adjective (I)
Outmaneuver	[out [maneuver]]	Root: maneuver = Verb out + Maneuver => Verb (D)
Actually	[[actual] ly]	Root: Actual = Adjective Actual + ly => Adverb (D)
Moved	[[move] ed]	Root: Move = Verb Move + ed => Verb (I)
Almost	[al [most]]	Root: Most = Adjective al + Most => Adverb (D)
Surrounding	[[sur [round]] ing]	Root: Round = Adjective Sur + Round => Verb (D) Verb + ing => Adjective (I)
Favorite	[[favor] ite]	Root: Favor = Noun Favor + ite => Noun (D)
Becomes	[[become] s]	Root: Become = Verb Become + s => Verb (I)

It becomes clear from the data in this table that in spontaneously spoken Anglo-Estonian dialogue, there is a clear lack of complex polymorphemic words. Instead, an abundance of simpler bimorphemic words is found, with the collection of trimorphemics present in the corpus numbering only a handful. Only one polymorphemic word was found with more than three constituent morphemes,

which would seem to suggest a tendency in spontaneous Anglo-Estonian dialogue towards shorter word encodings.

SYNTACTIC ANALYSIS

In a continuation of the morphological examination performed above, a detailed assessment of syntactic usage was also executed. Traditional phrase structure rules were applied to five randomly selected sentences and phrase diagrams were generated for each. Diagrams I-V below include the source sentence, the phrase diagram, and the lexicon fragments representing the subcategorizations of each component verb. The following encoding scheme was used for part-of-speech tokenization:

S = Sentence

NP = Noun Phrase

VP = Verb Phrase

DET = Determiner

CC = Coordinating Conjunction

ADJ = Adjective

ADV= Adverb

PREP = Preposition

PP / P = Prepositional Phrase

AUX = Auxiliary

PRON = Pronoun

DIAGRAM I

Sentence	These shows were actually very clever and were not demeaning to these hill billy family.
Phrase Diagram	S => NP => DET = These Noun = shows VP => V = were ADV = actually NP => ADJ = very Noun = clever CC => and VP => V = were ADV = not ADV = demeaning PP => PREP = to NP => DET = these ADJ = hill ADJ = billy N = family
Lexicon Fragments	Were: Verb, V _____ ADV

DIAGRAM II

Sentence	The early movie, early shows, from around 1964 till about 66 were black and white.
Phrase Diagram	S => NP => DET = The ADJ = early Noun = movie NP => ADJ = early Noun = shows PP => PREP = from PREP = around NP => Noun = 1964 PP => PREP = till PREP = about NP => Noun = 66 VP => V = were NP => ADJ = black CC = and ADJ = white
Lexicon Fragments	Were: Verb, V _____ ADV

DIAGRAM III

Sentence	That may be even more stupid, but there was a lot of truth, at least in the first show.
Phrase Diagram	S => NP => DET = That VP => AUX = may V = be ADV = even NP => ADJ = more N = stupid CC => but NP => DET = there VP => V = was PP => PREP = a NP => ADJ = lot PP => PREP = of NP => N = truth PP => PROP = at NP => ADJ = least PP => PROP = in NP => DET = the ADJ = first N = show
Lexicon Fragments	Be: Verb, V _____ ADV Was: Verb, V _____ DET

DIAGRAM IV

Sentence	And the, my favorite show is the Beverly Hill Billies
Phrase Diagram	S => CC => and NP => DET = the NP => PRON = my ADJ = favorite N = show VP => V = is NP => DET = the ADJ = Beverly ADJ = Hill N = Billies
Lexicon Fragments	Is: Verb, V _____ NP

DIAGRAM V

Sentence	He comes in and they start negotiating to get property.
Phrase Diagram	S => NP => PRON = He VP => V = comes PP => PREP = in CC => and NP => PRON = they VP => V = start V = negotiating PP => PREP = to VP => V = get NP => N = property
Lexicon Fragments	Comes: Verb, V _____ PP Start: Verb, V _____ V Negotiating: Verb, V _____ PP Get: Verb, V _____ NP

The results of the syntactic analysis of these five sentences coincides directly with the conclusion exhibited by the morphological analysis of the corpus: the spontaneity of Anglo-Estonian speech favors short encodings. Where the morphological decomposition yields short word encodings, the syntactical assessment demonstrates short thought encodings, with frequent incoherence caused by byzantine intersentential referencing and recurrent thought occlusion. Such incoherence required the introduction of Treebank Coordinating Conjunction branches into the phrase diagrams to properly delineate sentential flow sequences. These branches are prefaced by a “CC” symbol, followed by the

word the break occurred on. The lack of lucidity and direct translation between sentence fragments necessitated the use of TCC's, rather than standard sentential linkages. Additionally, several other mutations were required of the standard English phrase structure rules, including the generation of the following structures:

- CPV (Complex Verb Phrase): capable of encoding sequential verb representations
- CPP (Complex Prepositional Phrase): capable of encoding sequential prepositions
- CNP (Complex Noun Phrase): capable of encoding multiple first-order representations (arbitrary entity class)

These structures are included in the phrase diagrams as simply "CP", "PP," and "NP," respectively. Their generation was indicated by several grammatical flaws in the selected corpus, which could not otherwise be directly resolved or reconciled with standard phrase structure rules.

SEMANTIC / PRAGMATIC ANALYSIS

As the final stage of orthographic analysis, the semantic and pragmatic properties of the transcribed Anglo-Estonian speech were analyzed. A selection of ten continuous sentences was chosen at random from the global corpus and the thematic role of each verb was delineated. At the same time, intersentential referencing was examined for each sentence, and a summary of the results are presented below in Diagrams VI-XVI. For intersentential referencing, a

designation of “Global External Referent” signifies a referent which is not directly present in the document, such as the audience or speaker, while “External Referent” simply represents an entity which was introduced in an earlier sentence, and an “Internal Referent” is an entity which was introduced in the same sentence. Duplicate entries in the Thematic Role and Intersentential Reference fields have been removed.

DIAGRAM VI

Sentence	You might find that absurd and stupid, and if you looked at it recently you're probably right, if you've seen it on Nick on TV, but if look at it from the timeframe of the 1960's and you look at a time frame of the early shows, not the late shows, there's rather good and and fun humor in it.
Thematic Roles	Find, V, _____ PP, (Agent, Theme) Look, V, _____ PP, (Agent, Theme) Are, V, _____ NP, (Agent, Theme) Have, V, _____ V, (Agent, Theme) Seen, V, _____ NP, (Agent, Theme)
Intersentential References	You = Audience (Global External Referent) That = Beverly Hill Billies (External Referent) It = Beverly Hill Billies (External Referent)

DIAGRAM VII

Sentence	They moved to California, the Clampets do, and the first couple years with the black and whites, they were clever, people were trying to take their money away from them, and they always seemed to manage to keep the money going and so its its a show that shows that even though you're not educated, you have a really good chance of surviving.
Thematic Roles	Move, V, _____ PP, (Theme, Goal) Do, V, _____ (Agent) Were, V, _____ NP, (Agent, Theme) Seemed, V, _____ PP, (Experiencer, Theme) Going, V, _____ (Theme) Is, V, _____ NP, (Agent, Theme) Shows, V, _____ PP, (Agent, Theme) Have, V, _____ NP, (Agent, Theme) Surviving, V, _____ (Agent)
Intersentential References	They = Clampets (Internal Referent) You = Audience (Global External Referent)

DIAGRAM VIII

Sentence	In the first show, the Jed Clampet knows about this oil, the trailer that always goes in front shows Jed Clamplet shooting the oil and running around saying "we're rich we're rich we're rich".
Thematic Roles	Know, V, _____ PP, (Experiencer, Theme) Goes, V, _____ PP, (Theme, Goal) Shows, V, _____ NP, (Agent, Theme) Are, V, _____ NP, (Agent, Theme)
Intersentential References	We = Clampets (External Referent)

DIAGRAM VIV

Sentence	But, in reality, what happened was there a geologist who was prospecting in St. Clair county, Missouri, according to the literature on this show and he found an oil show.
Thematic Roles	Happened, V, _____ V, (Agent, Theme) Was, V, _____ NP, (Agent, Theme) Prospecting, V, _____ PP, (Agent, Location) Found, V, _____ NP, (Agent, Theme)
Intersentential References	Who = geologist (Internal Referent) He = geologist (Internal Referent)

DIAGRAM X

Sentence	Ok you asked me to give a talk on something and talk about my favorite TV show is what I'm gonna do.
Thematic Roles	Asked, V, _____ NP, (Agent, Theme) Give, V, _____ PP, (Agent, Theme) Talk, V, _____ PP, (Agent, Theme) Going, V, _____ PP, (Theme, Goal) Do, V, _____ (Agent)
Intersentential References	You = Audience (Global External Referent) My = Speaker (Global External Referent) I = Speaker (Global External Referent)

DIAGRAM XI

Sentence	Its really based on early show like the movie Sergeant York with Gary Cooper in that you have a hill billy who comes in and becomes the most highly decorated officer of World War I
Thematic Roles	Based, V, _____ PP, (Agent, Theme) Have, V, _____ NP, (Agent, Theme) Comes, V, _____ PP, (Agent, Goal) Becomes, V, _____ NP, (Theme, Goal)
Intersentential References	It = Beverly Hill Billies (External Referent) You = Audience (Global External Referent) Who = hill billy (Internal Referent)

DIAGRAM XII

Sentence	And you think of Sergeant York, which is based a true story, you don't find anything absurd with it, but when you go and look at the Beverley Hill Billies, you say "oh that's stupid", but its not, the Beverly Hill Billies is of course about a family, a hill billy family, that finds, somebody finds oil on their property and they move to the Beverly Hills of California to live the rich style.
Thematic Roles	Think, V, _____ PP, (Agent, Theme) Is, V, _____ V, (Agent, Theme) Based, V, _____ NP, (Agent, Theme) Do, V, _____ V, (Agent, Theme) Find, V, _____ NP, (Agent, Theme) Go, V, _____ (Agent) Look, V, _____ PP (Agent, Theme) Say, V, _____ NP (Agent, Theme) Move, V, _____ PP, (Theme, Goal) Live, V, _____ NP (Agent, Theme)
Intersentential References	You = Audience (Global External Referent) It = Sergeant York (Internal Referent) It = Beverly Hill Billies (Internal Referent) Their = hill billy family (Internal Referent) They = hill billy family (Internal Referent)

DIAGRAM XIII

Sentence	The geologist uh was working for OK Oil Company in Tulsa Oklahoma, which is where the oil business used to be uh based at before almost all of it moved to Houston.
Thematic Roles	Was, V, _____ V, (Agent, Theme) Is, V, _____ V, (Agent, Theme) Used, V, _____ V, (Agent, Theme) Based, V, _____ PP, (Agent, Theme) Moved, V, _____ PP, (Theme, Goal)
Intersentential References	It = oil business (Internal Referent)

DIAGRAM XIV

Sentence	You had Jed Clampet, Grampy, Ellie May, and a number of other characters, and that Jed Clampet was always in these early shows very uneducated, did not know anything really about modern life, but had a common street smart sense and was able to outmaneuver, just by being smart, all these highly educated people.
Thematic Roles	Had, V, _____ NP, (Agent, Theme) Was, V, _____ PP, (Agent, Theme) Know, V, _____ PP, (Experiencer, Theme) Outmaneuver, V, _____ (Agent)
Intersentential References	You = Audience (Global External Referent) These = highly educated people (Internal Referent)

DIAGRAM XV

Sentence	One of the reasons I liked the Beverly Hill Billies is that it, was at least in the pilot episode, for a geologist, it was really a fun uh thing show to watch, the pilot was was a geologist was in the pilot and he went and he talked about finding oil on the Clampet ranch, the Clampet uh property in Bug Hollow Missouri.
Thematic Roles	Liked, V, _____ PP, (Agent, Theme) Is, V, _____ PP, (Agent, Theme) Was, V, _____ PP, (Agent, Theme) Watch, V, _____ (Agent) Went, V, _____ (Agent) Talked, V, _____ PP (Agent, Theme)
Intersentential References	I = Speaker (Global External Referent) It = Beverly Hill Billies (Internal Referent) He = geologist (Internal Referent)

The semantic and pragmatic analyses of the orthographic transcription yield several interesting trends. The first is a sparing use of intersentential references. While many native English speakers tend to over-emphasize their use of such referencing, the subject uses them quite sparingly, decreasing the load on short-term memory usage for transference storage. However, a marked preference for discordance with English grammatical rules was observed in the thematic role analysis of the text. The roles stated above were taken from the verbs' usage within the text, rather than their classical models. A substantial deviation is noted in several of the verbs, which contribute heavily to overall incoherence though many of the thought clusters. These deviations made it difficult to identify the full thematic roles of some verb instances.

PHONETIC / PHONOLOGICAL ANALYSIS

Upon completion of the orthographic portions of the analysis, phonological exploration was embarked upon, to document substantial features in the spoken Ango-Estonian word. Three sentences were selected at random and phonetically encoded in IPA (International Phonetic Alphabet) format. The orthographic transcription and IPA phonetic encoding of each sentence appears below in Diagrams XVI-XVIII, as well a discussion of any unusual characteristics found in that sentence.

DIAGRAM XVI

Orthographic Transcription	And the, my favorite show is the Beverly Hill Billies.
Phonetic Encoding	ᵊn ði maɪ fevɜrɪt ʃo ɪs ðᵊ bɛvɜrli hɪl bɪlis
Commentary	The pronunciation of the first word, “and,” differs in its ending from the standard isolation pronunciation. The “d” sound at the end of the word was dropped to ease the transition to the next word, “the”, which carries the same phonetic encoding as the deleted ending of “and.”

DIAGRAM XVII

Orthographic Transcription	The early movie, early shows, from around 1964 till about 66 were black and white.
Phonetic Encoding	ðə ɚrli mu ɚrli ʃos fr ʌ m æran nɪtinsɪkstifor tɪl baut sɪkstisɪks wɛr blæknwaɪt
Commentary	Although the orthographic transcription shows the word "movie" in its entirety, the subject failed to complete this word utterance, opting instead to backtrack one word and repeat his original thought, "early shows." Since backtracking is present in this sentence, that indicates a reevaluation of the deep store memory image used to construct the sentence, which in turn would suggest a sound environment dependant on the precursory word "early." The first sound in "about" is truncated from its pronunciation in isolation. This is caused by the acoustic environment installed by the word "till" that appears immediately before it. After the pronunciation of "till," very little additional preparations must be made for the word "bout" to be uttered, whereas "about" would require additional sound manufacture. The "and" between "black and white" is truncated to a simple "n" sound. This is most likely an effect of the neutrality of the "n" sound, which does not require substantial alteration of the sound environment for production, whereas "and" would require an entirely new environment for its production.

DIAGRAM XVIII

Orthographic Transcription	These shows sh were actually very clever and were not demeaning to these hill billy family.
Phonetic Encoding	ðiz ʃos ʃ wɛr ækʃli veri klɛver ən wɛr nat dminin tu ðiz hɪl bɪli fæmli
Commentary	An additional "sh" sound was introduced after the word "shows," which is indicative of a "slip of the tongue" mispronunciation. This is most likely caused by the fact that an "sh" both precedes and succeeds the noun in question, which could cause the tongue to over-anticipate and place itself into position for the next utterance, which failed to be produced.

The deviations from standard found in the subject's pronunciation follow closely the spontaneity degradations observed in the orthographic analyses. As the level of expressed thought increases, the sentential coherence and grammatical validity decreases. However, the phonetic encodings also reveal the influence of regional sociolinguistic influences on grammar constructs and individual word pronunciations, which could account for some of the irregularities.

DISCUSSION / CONCLUSIONS

From the selected orthographic and phonetic analyses of the subject's speech sample, several marked trends may be observed. The first of those is the selection of morphologically simple words, which lessens the burden of stem translation and frees the short-term memory cortex for other tasks. The second is a substantial departure from the classical usage models of many common verbs and grammatical constructs. These elements were put together in novel arrangements during periods of apparent significant activity in the short-term memory areas, suggesting that the incoherence and thought occlusion is caused by Cortex Access Deprival (CAD), which is a common occurrence in computational models that attempt to approximate language dialects. Future exploration could include dynamically-altered queries designed to produce and exploit CAD environments in an attempt to replicate such behavior in a stable form.

Overall, the subject's analyzed speech proved to be more complex than classical examples provided in many texts, as the particular nature of the grammatical

constraint violations present in the selected corpus were difficult to replicate within other collections. This required the creation of several new phrase structure rules to accommodate sentential incoherence and construct misuse, which were not addressed in the source texts with which the author was familiar. In addition, thematic roles were not easily identified, due to the complexity and novelty of many of the constituent constructs.

Nevertheless, despite these limitations, a classical quantification of the Anglo-Estonian dialect has been presented through the incorporation of stochastic language process techniques. This comprehensive profile should be used as the basis for future research into the concerns addressed herein.

APPENDIX I

Ok you asked me to give a talk on something and talk about my favorite TV show is what I'm gonna do. And the, my favorite show is the Beverly Hill Billies. You might find that absurd and stupid, and if you looked at it recently you're probably right, if you've seen it on Nick on TV, but if look at it from the timeframe of the 1960's and you look at a time frame of the early shows, not the late shows, there's rather good and and fun humor in it. Its really based on early show like the movie Sergeant York with Gary Cooper in that you have a hill billy who comes in and becomes the most highly decorated officer of World War I. It was made in World War II as a propaganda movie, but the premise is exactly the same as the Beverly Hill Billies. And you think of Sergeant York, which is based a true story, you don't find anything absurd with it, but when you go and look at the Beverley Hill Billies, you say "oh that's stupid", but its not, the Beverly Hill Billies is of course about a family, a hill billy family, that finds, somebody finds oil on their property and they move to the Beverly Hills of California to live the rich style. The early movie, early shows, from around 1964 till about 66 were black and white. These shows sh were actually very clever and were not demeaning to these hill billy family. You had Jed Clampet, Grampy, Ellie May, and a number of other characters, and that Jed Clampet was always in these early shows very uneducated, did not know anything really about modern life, but had a common street smart sense and was able to outmaneuver, just by being smart, all these highly educated people. One of the reasons I liked the Beverly Hill Billies is that it, was at least in the pilot episode, for a geologist, it was really a fun uh thing show to watch, the pilot was was a geologist was in the pilot and he went and he

talked about finding oil on the Clampet ranch, the Clampet uh property in Bug Hollow Missouri. They don't talk about Missouri, but the original writer was from Missouri, and from St. Clair county, which is south of Kansas City, and as I as I teach a graduate level course in petroleum geology, I sometimes use the Beverly Hill Billies as a standard of what the oil industry is like. That may be even more stupid, but there was a lot of truth, at least in the first show. In the first show, the Jed Clampet knows about this oil, the trailer that always goes in front shows Jed Clampet shooting the oil and running around saying "we're rich we're rich we're rich". But, in reality, what happened was there a geologist who was prospecting in St. Clair county, Missouri, according to the literature on this show and he found an oil show. Now an oil show is actually used as prospecting, is not that common, because most oil shows have now been found, but at that, the premise was that there was an oil show, the Colonel Drakes Well in Pennsylvania, they found and drilled the first commercial oil well in the United States because of an oil seep into the surrounding stream. And so that's true, he comes in, then it gets a little bit absurd, Ellie May hits him, but the reason he hits him, he thinks, she thinks he's a revenuer, well a revenuer wants to go after granny's sill, that's a legitimate thing, in fact, that's what they would do, they'd probably shoot the guy and ask questions later, because they're worried about people going and getting rid of some of their home brew. Then, as he comes back and he becomes a land man and a geologist, and if you talk to some of the independent operators oil operators in this basin, that's what they are too, they are a geologist, but then they're also negotiators with the land owners, in this case, the Clampets. He comes in and they start negotiating to get property. The thing was that the oil show was so huge that the Clampets said "ah, they want twenty-five to a hundred

dollars". Hundred what? "It's a new type of money, millions". And suddenly, Jed Clampet and his family were millionaires. The geologist uh was working for OK Oil Company in Tulsa Oklahoma, which is where the oil business used to be uh based at before almost all of it moved to Houston. They moved to California, the Clampets do, and the first couple years with the black and whites, they were clever, people were trying to take their money away from them, and they always seemed to manage to keep the money going and so its it's a show that shows that even though you're not educated, you have a really good chance of surviving.