Hypocoristics:

A derivational Problem

Mohammed Al-Qenae

Master of Arts by Research

University of York

Language and Linguistic Science

May 2018

Abstract

This study is an investigatory research on the two major schools of linguistics, formal and functional. The study looks at earlier versions of Generative Theory as the representative of formal linguistics and contrasts it to Skousen's computational model which is taken as the representative of functional linguistics. The way each of the theories are described and evaluated are by considering how each of them can be used in analysing hypocoristic data. A description of hypocoristics for 165 names collected from Kuwaiti Arabic speakers were the base for the analysis. The data was given a general description at first to show how they can be accounted for in the two theories. The first approach that was used was a rule-based approach used previously with Jordanian Arabic Hypocoristics which use Semitic root and Pattern Morphology. The second rule-based approach was also a rule-based approach the employed phonological processes to account for the derivation. The two were considered part of formal theories of analysis. The functional analysis which uses a computational model that employs phonological features defined over statistically driven frequencies was used to model the data. An evaluation of the model with low success rates lead to the change of the model and present an alternative hybrid model that utilises both rules and analogy. The model was inspired by a rule-based theory which was not fleshed out and analogy was used to flesh it out and place it with a usage-based theory of language. Finally, the thesis ended with an open evaluative stand requiring further research on computational models from a computational perspective rather than a linguistics view.

Table of Contents

List of Tables

List of Figures

Dedication

To Hadduy.

Acknowledgment

I would like to thank my advisor Dr. Sam Hellmuth, for everything, from accepting me as a student, to showing me both how the academic field works, and to all the support that she gave me. I would like to especially thank her for her patience and tolerance which I believe no other supervisor can provide. I will forever hold dearly what she has shown me through my time at the University especially her concern for my family's wellbeing.

I would also like to thank, Dr. Amin Al-Muhanna and Prof. Jean François Prunet. Both of them were there when I had questions regarding any academic and linguistic details. Without Professor Prunet and the many correspondences this project would have never been completed. I would like to thank him for giving me the time and guidance which he was not obliged to commit to. I would also like to that from the University of York, the Maths Skills Centre for the support that was received in analysing the data. I would like to thank Professor Dunstan Brown, Professor Peter Sells, and Márton Sóskuthy.

I would also like to thank my parents. I thank them for being there by choice and not by force. I would like to thank them for the support they have shown both me and my wife and thank them for the laughter they have brought to us. I will always cherish what they have shown me and taught me.

There is one final person that no words will ever be enough to express the gratitude I hold towards them- my dear wife. She endured the most out of this experience by putting up with my ups and downs. She truly must have all of the regards and esteem that this work as well as any other future endeavours will have. She was my sanity throughout the many hardships. Without her this work would not have been possible. Without her I wouldn't have survived.

Declaration

I undertake that all the material presented for examination is my own work and has not been written for me, in whole or in part, by any other person. I also undertake that any quotation or paraphrase from the published or unpublished work of another person has been duly acknowledged in the work which I present for examination. This work has not previously been presented for an award at this, or any other, University.

1. Hypocoristics: A Derivational Problem

In the first sentence of a paper titled The Many Faces of Nicknames, the author states that "Despite a vast scholarly literature on general naming practices, the serious investigation of nicknaming has barely begun" (Holland, 1990, p. 255). The statement is in conflict with a concluding remark from a recent paper that discussed Arabic nicknames, where the authors comment that "no linguist can speak for the community when it comes to the analysis of names, since we are dealing with extralinguistic competence" (Idrissi, Prunet, & Béland, 2008, p. 247) (Henceforth IPB). The categorisation of nicknames as being extralinguistic is mainly due to having speakers' usage of hypocoristic exhibit a "total freedom of analysis" (Idrissi et al., 2008, p. 250). This is not the only time that nicknames have been considered an unanalysable phenomenon, irregular, marginal, or an oddity (Aronoff, 1976; Dressler & Barbaresi, 1994; Stonham, 1994; Zwicky & Pullum, 1987). This raises questions on the validity of the claim that nicknames are not part of grammar. If true, it dissuades linguistic investigation on nickname analysis.

In this thesis, one main question will be answered: are hypocoristics irregular and as a result should they be considered extra-grammatical? In other words, can a hypocoristic analysis be accomplished or should they be deemed an unanalysable phenomenon similar to how language games were considered prior to Prosodic Morphology (Bagemihl, 1995). The answer to the question, and a summary of the thesis is **that hypocoristics do show unpredictable patterns and depending on the theoretical approach taken these patterns can be reduced leading to a conclusion that hypocoristics are not extra-grammatical and a descriptive analysis is possible.**

The answer will be given by providing two different dominant theories of analysis; a rule-based analysis that was previously used with Arabic Nicknames and an analogy-based analysis that is novel to Arabic hypocoristic investigation. A description of Arabic nicknames for 165 names, collected from Kuwaitis, will be the base for the analysis. It will be given in the second chapter using a very general approach of description whereby the two analyses or any future analysis can use it. The description of the data will be provided along with the way it was collected.

Following the description of the collecting approach and the data, a rule-based analysis will be presented in chapter three. There are two main rule-based approaches that have been previously used with another data set on Arabic Hypocoristics. The two approaches will be explained and will be applied to the collected Kuwaiti Arabic data set. The chapter will end by comparing the two rule-based approaches and evaluating their success in accounting for the data using a simple equation.

Chapter four will start with defining analogy as it was used in other studies and how it will be used in the present study. The way it will be used here is through a pre-existing computational algorithm, AM, which models the data. The algorithm will be explained before presenting the results of the model which will be discussed briefly before taking into consideration other theoretical issues treated in the following chapter.

In final chapter both rule-based analyses and the analogy analysis will be discussed by comparing them on different theoretical issues. The main issue is provided mostly by IPB’s claims that semantics of proper names are problematic for any analysis that employs semantic decomposition. There are other issues that have been raised and they will all be used to discuss and evaluate the different given analysis. The chapter will not suggest one approach being superior to another. Instead it will conclude that a third hybrid approach can be drawn between the two approaches and can be employed. The hybrid approach will be used to analyse the main part of the data. The findings given from the linguistic analysis and discussion will surely benefit “the serious investigation of nicknaming” (Holland, 1990, p. 255).

1.1 Questions & Terminology

Some of the topics that will be discussed later in detail require a brief introduction and some definitions. One of the main reasons that these topics need to be explained here is due to having more clarity required as these terms have in the literature different interpretations. The explanation given here will be the ones used in the thesis.

What is Cognitive Linguistics?

The answer to this question should be, ‘which cognitive linguistics?’. Cognitive Linguistics is different than cognitive linguistics. Geeraerts (2006) explains “terminologically […] we now need to make a distinction between Cognitive Linguistics […] and uncapitalized cognitive linguistics – referring to all approaches in which natural language is studied as a mental phenomenon. Cognitive Linguistics is but one form of cognitive linguistics, to be distinguished from, for instance, generative grammar and many other forms of linguistic research within the field of cognitive science.” (p. 3).

Geeraerts (2006) then questions the specificity of Cognitive Linguistics since it is not a single platform but rather a research paradigm. He best uses an analogy in describing it as “an archipelago rather than an island. [where] It is not one clearly delimited large territory, but rather a conglomerate of more or less extensive, more or less active centers of linguistic research that are closely knit together by a shared perspective, but that are not (yet) brought together under the common rule of a well-defined theory.” (p. 2).

Geeraerts (2006) then shows the various works and instantiations that has been labelled Cognitive Linguistics and represent the archipelago. In this thesis, the specific instantiation of Cognitive Linguistics that will be discussed is Usage-based Theory. Although it is also referenced in the works of Bybee (2001, 2007, 2010), Pierrehumbert (2001), and Tomasello (2005). This particular instantiation follows the work of Skousen and his successors (Chandler, 2009; Eddington, 2000, 2004, 2009; Skousen, 1989, 1992; Royal Skousen, Deryle Lonsdale,

& Dilworth B. Parkinson, 2002a). Thus, what should be taken is that the Usage-based analysis given in this thesis might have a lot of similarities with Cognitive Linguistics and Usage-based Theory as discussed in the literature; here it is mostly restricted to Skousen's work. Skousen's work will be referred to in this thesis as usage-based theory, Cognitive Linguistics, analogy-based theory, and functional linguistics.

What is Generative Grammar?

The restriction of Cognitive Grammar to Skousen's work has a parallel in the approach which will also be used with Generative Theory. Generative Grammar has also seen changes to the theory and is referenced with various instantiations of it. Since Skousen's work has been directed against rule-based grammar as set out in Chomsky and Halle's Sound Pattern of English and Chomsky's Aspects of the Theory of Syntax, it is mainly the work in these two that will be taken as the representative of Generative grammar[1]. In this thesis, Generative Theory will be related to as Chomskyan Grammar, generative grammar, rule-based theory, and formal linguistics.

What is Grammar?

The two approaches both deal with the description of grammar. The term grammar is used in linguistics in more than one way. Along with mathematics and astronomy, grammar constituted a science of language that is part of Philosophy (Robins, 1997). The Greeks used it as the study of writing and reading. It was then developed by the Romans as the rules that constituted what was correct in a language. Modern linguistics continued such a usage of grammar as rules of a language. However, they divided grammar into prescriptive and descriptive. Like the Roman use of grammar, prescriptive grammar acts as an established

[1] Even though changes to the theory has been proposed by both Chomsky and his successors and even though Structuralists have used rules for their grammar, Chomsky and Halle (1968) and Chomsky (1965) are still widely regarded as the opposing approach to Cognitive Linguistics and the representative of rule-based grammars (Guy, 2014). Thus, this restriction is widely accepted.

constitution that imposes what is right and wrong in a language from its meaning to the way it sounds.

Where modern linguists deviated from the traditional grammar, is with ending a long history of language study that imposes rules upon speakers. Instead, a grammar is seen as a complex system that underlies what the speaker knows- language competence. It is the description of the judgement and intuition of native speakers. Thus, it places the power of establishing what is right or wrong in a language with native speaker's competence[2]. Furthermore, descriptive grammar still maintained being of a rule format with Generativists where it makes up part of a system that is responsible for speech[3].

The topic of prescriptive grammar versus descriptive grammar is one that is discussed in almost every linguistic introduction textbook. The importance of it here is not the distinction between the two but the relation that prescriptive grammar has with descriptive grammar as set by functional and formal linguists. With formal linguists, one can connect a line between the grammar they form and the one that is being taught as a prescriptive grammar since both use rules and conditions as a formula (F. Newmeyer, 1988). With functional linguists, the description of analogy for example, is not and should not be formed with rules. This leaves a functional descriptive grammar not easy to describe and present. This complication will be demonstrated in the thesis in that the description of hypocoristic formation will easily be shown in the form of grammar rules in chapter three but will be explained as an abstract theoretical process in chapter four.

[2] Both formalists and functionalists agree on having competence or the mentalist knowledge of the speaker as the source of grammar but they do not agree on the role of performance or language use. Formalists exclude it from the grammar while functionalists include it (Bybee, 2001, 2007). Note that this point has been discussed in usage-based theory literature where some do have the distinction between performance and competence but with Skousen and his successors the distinction is questioned (Eddington, 2004).

[3] The system as a whole can be called grammar which can be divided into two components, a lexicon and a grammar which has the rules of a language.

What is extra-grammatical?

For generativists, the rules in a grammar describe the parts of the language that do show regular and formal pattern. This leaves a set of language that exhibits irregularity whereby a rule cannot be used to describe them. This irregular set is said to be extra-grammatical. (Dressler & Barbaresi, 1994). This invokes another closely related term, extra-linguistic, which is what accounts for the production of the irregular forms. Formations that are not based on rules, but instead are achieved by external factors are said to be extra-linguistic. Acronyms, reduplicatives, truncations, and analogy formations are all considered to be extra-grammatical and acquired through extra-linguistic factors rather than an innate rule-system (E. a. Mattiello, 2013).

For functionalists on the other hand, “No relevant methods for gaining evidence about language are excluded” (Heine, Narrog, Bybee, & Beckner, 2009, p. 827). Diachronic and synchronic data, corpus studies, social variation findings, and performance errors are all valid sources for gaining evidence. On the other hand, generativists believe these findings with the exclusion of synchronic data, to be part of extra-grammatical formations. Thus, the two terms, extra-grammatical, and extralinguistic are more restricted to Generative Grammar.

What is Productivity?

For Generativists, one tool that can be used to determine what is part of grammar is productivity. The topic of productivity is one of the most debated topics in linguistics (Bauer, 2001, 2005; Lieber, Štekauer, Aronoff, & Lindsay; Rainer, 2005). The result of the debate is not having a single way of defining it and many approaches are taken in defining productivity (Bauer, 2005; Scherer, 2015).

The best way to explain productivity in terms of how it will be used here, is by looking at Diachronic data. The English suffix *–th,* was used widely with adjectives to form nouns as in the words *depth*, *strength*, *length*, and *truth*. The morphological process today is no longer used. As a result, linguists would conclude that *–th* affixation was a productive process in Old English and is not productive today.

A quantitative formalism that is used to determine whether a process is productive or not is by looking at the frequency of its applicability[4]. To do that, a count of tokens can be taken and/or a count of types. In this thesis, it is the frequency of type that will be used as an indicative of productivity- as explained below.

What are Types & Tokens?

There are two types of items that play a role in a quantitative measure of productivity, type and token[5]. Types are the unique items that are found in a data-set excluding repetitions of that same unique item. On the other hand, tokens are the total of every item including repeated ones of the same item. Table 1 illustrates the difference between the two:

Table 1

The difference between type and token as used with hypocoristics.

	Name	Truncation	Trunc + -y/ie Affix
a.	Samantha	Sam (9)	Sammie (5)
b.	Christine	Chris (4)	Chrissy (1)
c.	Joseph	Joe (2)	Joey (3)
d.	Dorothy	*Dor (0)	Dory (5)
		Type: 3 Token 15	**Type: 4 Token: 14**

In Table 1, examples of nicknames are given to illustrate the difference between type and token. There are two nickname formation processes used in English. The first is truncation of the name to the left-most CVC syllable and the second is truncation plus a *y/ie* suffix. If productivity of a process is calculated based on type then a look at the number of names the pattern has been used with without repetitions is what should be looked at. In other words, how many names can be truncated to form a nickname. The answer will be three as it applies only

[4] Potential applicability will also be used in this study as a way of testing a theory.

[5] A third item, hapax legomena or words occurring only once in a data-set, is also used but won't be taken here.

to the first three names[6]. On the other hand, if productivity was a process that is based on token, then a count of how many times it applied should be taken. It will include a count of all repetitions and non-repetitions. In Table 1, it will be understood that truncation was applied in the data by subjects 15 times and 14 times with the truncation along with the *-y/ie* Affix.

What is Onomastics?

Onomastics is the study of proper names. The work done in onomastics is concerned with corpora studies, diachronic work, etymology finding, and social details of names all of which are marginal areas for Generativists. The peripheral status of these areas given by the dominant study of linguistics from a Generative perspective may have led to separating onomastics from Linguistics where once they were considered a single field of language study[7] (Hajdú, 2003). On the other hand, in a functionalist's perspective separation isn't required. Instead onomastics should be defined as the area of linguistics that is concerned with names rather than a separate field. This study is one which takes onomastics to be part of linguistics. In this study, a reoccurring theme will show how a division between the two fields is not warranted; starting from the next chapter where a description will be given.

[6] *Dot* is a nickname for *Dorothy* but it is questioned as to whether it is derived from *Dorothy* or whether it is associated with the name as part of the culture, similar to nickname *Bob* for *Robert.*

[7] No one has previously claimed that generative grammar is behind the division. However, this is assumed based on the dismissal of generative grammar objectives which dismisses the work and findings of onomastic research.

2. Names, Nicknames, and Hypocorism

It is usually the case that definitions of *proper names* given in dictionaries describe them as being nouns that are used to call a place, person, or an organization. Definitions might also mention having proper names be capitalized. These definitions are correct but in this thesis proper names will be limited to people's designators. Additionally, three other terms will be used to specify different types of proper names.

The three types of proper names can be established based on naming practices. When the term *name* is mentioned it will be reserved for the designator used for a person's birth name and official documents[8]. A name is the first given proper name that a new-born gets. On the opposite end, *nicknames* and *hypocoristics* are proper names that are given at any stage after a name is designated and are both not used for legal and official contexts (Brylla, 2016; Langendonck, 2007; Starks & Taylor-Leech, 2011). The difference between the two is that a hypocoristic has a phonetic resemblance to the name which is the result of a phonological or affixation process, while nicknames can be any other unofficial name. More precisely one can think of a nickname as a category which has under it various unofficial subordinate types of names like terms of endearment, derogatory names, patronyms, matronyms, teknonyms, and hypocoristics[9].

For example, on the fan page of one of the world's most famous soccer player, *David Beckham*, four nicknames are listed; *Dave*, *Becks*, *DB7*, and *Golden Balls*. Out of the four nicknames, only the first two are hypocoristics while the last two will only be called a nickname. The reason is that the first two have a phonetic resemblance with the name *David Beckham*. This raises the question on why *DB7* (initials and shirt number) is not a hypocoristic?

[8] The definition is general and might include last name, father name, or surname. The most important thing with this definition is having the term name be in opposite relation with nickname.

[9] The term byname is usually used instead of nickname as defined here. However, with most of the papers which are written in English and referenced here nickname is used and thus it will be adopted.

Although a phonetic resemblance is present with the initials, number 7 refers not to the name but some other external entity. Thus, it is not considered a hypocoristic. Not only that but its unproductivity is seen with it being restricted to *Beckham*.

As will be later discussed, phonetic resemblance to the name is the main feature which differentiates hypocoristics from other nicknames, but sometimes it is not enough since no productive phonological process can be seen applied to the name. For that reason, a productive phonological process is the main feature required for a nickname to be called a hypocoristic. Having it be a productive process specifies hypocoristics as a synchronic phenomenon. This will diverge and leave out many onomastic usage of the term hypocoristics when referring to a name that etymologically was produced through a phonological process but is no longer used (Brylla, 2016; Ion, n.d.). In short, the name and nickname/hypocoristic distinction falls in a precedence naming relation where the task of nicknaming can't be performed unless a name exists; and out of the various nicknames, hypocoristics are formed using a productive phonological process that is applied to the name and can be described.

2.1 Nicknames & Hypocoristics

The example of Beckham's four nicknames is not something unusual, nor is it something common only among sports players (Kennedy & Zamuner, 2006; Taylor & Kennedy, 2015). When looking at nickname descriptions in other languages one can see that the number of nicknames a single person can have reaches up to sixteen and some of them are hypocoristics. This is witnessed since the Greeks with names such as *Alexandros* which has *Alex*, *Alexis*, *Alekos*, and *Alekakos* as its hypocoristic (Leibring, 2016).

In some cases, the hypocoristics are very similar to each other and treating them as distinct hypocoristic formations or not is a question that has been tackled in both linguistic and onomastic research. There are two main features that are used in order to treat two hypocoristics as two separate entities, formal and functional. A formal feature, takes an approach where if syntactic or phonological behaviour is different from one nickname to another then the two hypocoristics are different. For example, it is assumed that the hypocoristics of *Abe* and *Abbey* are formed by shortening *Abraham* to *Abe* and having an optional *–y* suffix. This is false on phonological basis. Formal evidence show that *Abbey* is not formed by having *Abe* be an "intermediate step in the derivation" (Lappe, 2007, p. 107). Abbey is derived separately making the two be treated as two separate hypocoristics. The functional features that are used to consider whether *Abe* and *Abbey* are one, employ semantic and social analysis. Since *Abbey* has a diminutive meaning and is used in different social contexts than *Abe,* then the two should be treated as two distinct hypocoristics.

The functional and formal features can be seen as minor features with the name *Abraham*. However, these features are important with Arabic language, should the hypocoristic *Badour* and *Badouri* for the name *Bader,* where the only difference is the suffix *-i,* be treated as different? Taking the formal condition, the two are the same where *Badour* can act as an intermediate step in the derivation. However, functionally they are different. As will be discussed in chapter 5, even though the names are hypocoristics with similar functional features, the suffix *-i* adds an additional meaning that results in difference in its usage. Thus, both will be treated here as distinct hypocoristics.

In addition to the two conditions acting as a segmental device, support for treating two similar nicknames as separate comes from subjects' intuitions; when asked to give nicknames for a single name, both *Badour* and *Badouri* were provided in various surveys entailing that they are considered different by subjects. This argument is important to ease up any claim that a language can have 16 different nicknames for one name[10]. Such a claim has definitely aided in the consideration of hypocoristics as an extra-grammatical phenomenon; since it is usually

[10] On a dedicated onomastic blog the name Elizabeth was mentioned to have 28 English nicknames. https://onomasticsoutsidethebox.wordpress.com/2017/02/06/the-many-nicknames-for-elizabeth/

the case that a grammatical derivation on a single domain has a single rule that can account for various items. However, by looking at nickname descriptions from other languages a multiple nickname feature should simply be looked at as a characteristic of nicknames and hypocoristics. In Table 2 a list is provided showing the number of nicknames that a single name can have as reported in 12 different languages[11].

Table 2:

Number of nickname patterns that a single name can have.

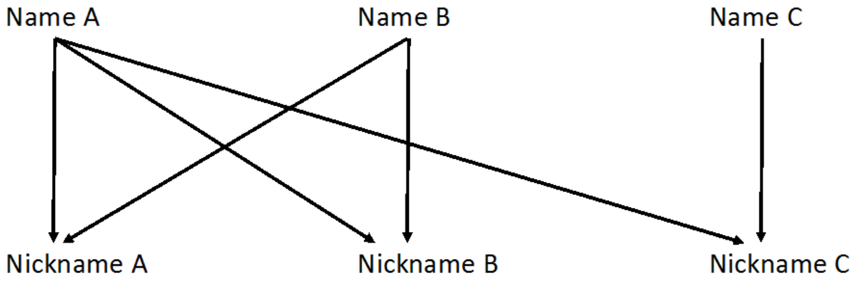
	Language	Number of Patterns	Reference
1	Indonesian	2	(Cohn, 2004)
2	Chinese	16	(Wang, 2004)
3	Bengali	6	(Lowe, 2004)
4	English	7	(Taylor & Kennedy, 2015)
5	Spanish	8	(Lipski, 1995)
6	Hebrew	3	(Bat-El, 2005)
7	Hausa	7	(Newman & Ahmad, 1992)
8	Mauritian	5	(Strandquist, 2004)
9	Basque	7	(Salaberri Zariategi, 2003)
10	French	3	(Plénat & Roché, 2001)
11	Japanese	3	(Mester, 1990)
12	Chamorro	2	(Robertson, 2004)

Another cross-linguistic feature related to having multiple nicknames and hypocoristics is the existence of cases where multiple names share the exact same nickname. For example, in English, the names *Samuel* and *Samantha* both have the hypocoristic *Sam*. This feature is common in Arabic and other languages. As a general depiction of the relation between names and nicknames one can form an image as shown in Figure 1.

[11] Even when such a broad categorization for nicknames is not taken, whereby *Abe* and *Abbey* are a single hypocoristic, the phenomena of having multiple hypocoristic patterns for a single name exists.

Figure 1:

Relationship between names and nicknames.



Before moving to the data used in this paper a summary of the two previous sections is required and given below:

- A proper name is a general category for any designator used to call a person.
- Names, nicknames, and hypocoristics are all types of a proper name.
- A name is the first given designator that a new-born receives and is used mostly in official and formal situation.
- A nickname is any proper name given at any stage after a person has a name.
- Nicknames are a general category that may include derogatory terms, pet names, patronyms, matronyms, teknonyms, and hypocoristics.
- Hypocoristics are the only nicknames that has a phonetic resemblance to the name due to a productive derivational process.
- Hypocoristics are phonetically similar to names because they are a type of nicknames which are formed using productive linguistic processes that apply to the name.
- A productive hypocoristic formation process can be based on a phonological process or a morphological concatenation process with the addition of affixes.
- A single name can have multiple nicknames and multiple hypocoristics.
- Two names can share the same nickname or same hypocoristic.

2.2 Data Collection

A list of Arabic nicknames was collected mostly from Kuwait University Students in the Arts College and Business Administration College. In this chapter, they will be presented by dividing them into nicknames and hypocoristics[12]. Note that they will be presented in order to accommodate the chapters that follow which will give an analysis on their formation and a detailed frequency of the collected item will be postponed to chapter four and five where the description will be relevant. Before presenting the data a description of the collection approach is required to better understand the data.

First Attempt

The first paper that attempted to analyse Arabic hypocoristic formation concluded that Arabic hypocoristics are unique (Davis & Zawaydeh, 1999) (Henceforth, D&Z). The paper had an analysis of hypocoristic data that was collected using a multiple-choice questionnaire where 11 native speakers of Jordanian Arabic had to choose from a number of acceptable and unacceptable hypocoristics. The subjects were also given the choice to provide another hypocoristic if none of the nicknames in the question seems acceptable.

Taking into consideration that there are multiple hypocoristics for a single name and different names for a single hypocoristic, a multiple-choice questionnaire will not be a good choice to elicit the available hypocoristics for a name. The method "would not give the subject freedom to elicit various existing patterns [...] [and that it] influenced the results [...] by provid[ing] [...] [more] guidance" (Idrissi et al., 2008, p. 85). With the objectives of both avoiding the weakness in D&Z's approach and questioning the uniqueness claim of Arabic hypocoristics a first collection attempt was set[13].

[12] The full list of names used for Hypocoristic analysis along with nicknames are in Appendix. They are organized using D&Z's approach which will be given in the next chapter.

[13] The first collection was made with the supervision of Jean-Francois Prunet who is one authors in the Idrissi et al. (2008) paper and went on to publish and present few other papers on Arabic Nicknames. He will be mentioned throughout the paper since his work had a strong influence on what is presented in this study.

A list of around 90 names was created. It consisted of a few names which D&Z used along with other random ones. There was no single collecting method taken. One-on-one recorded interviews were used, class discussions about hypocoristics and their meaning were conducted, and even different structure of questionnaire forms were distributed to students[14]. All of the approaches had weaknesses. However, they all contributed in the conclusion that D&Z are correct in claiming that Arabic Hypocoristics are unique. But the uniqueness hypothesis is not similar to D&Z's. Instead, the conclusion is that Arabic Hypocoristic formation has a lot of areas that require further investigation making it a good topic for research.

Second Attempt

A short analysis of the data from the first attempt was conducted which showed some hypocoristic patterns not discussed in D&Z's work. For that reason, two tasks were made. First a judgement task. Then, a data collection task was done for the unique names in addition to ones that had similar patterns. In the judgement tasks, students were asked to rank the acceptability of hypocoristic patterns that were provided by subjects from the first attempt. The acceptability was graded because in the first attempt it was noticed that acceptability of hypocoristics is gradient. Prunet worked on this topic using different approaches and showed that hypocoristic acceptance is gradient. He had a group of students answer questions about the acceptability of some nicknames. He then had another group invited and asked them about the same name while the first group were still seated. When there was a difference he turned the class into a discussion class leading all of the students to agree that the list of hypocoristic that they thought was unacceptable is acceptable (J. F. Prunet, 2014). Thus, the second attempt resulted in having the final list of names and finding the best possible approach to collect hypocoristics.

[14] In the first method family members and colleagues from work also participated.

Third Attempt

In the third attempt, 140 names were used to compile a hypocoristic collection that is analysed here. The names were divided across 4 sheets with 35 names each. Students were instructed to provide as many nicknames as they can think of for the given name. A total of 8 classes were used and each class filled out one of the four available sheets. A comparison of subjects' response to each survey showed a difference in the number of nicknames given. Below is a boxplot comparing the sheets, Figure 2 and Figure 3.

Figure 2:

Comparison of token frequency results taken from 4 different surveys.

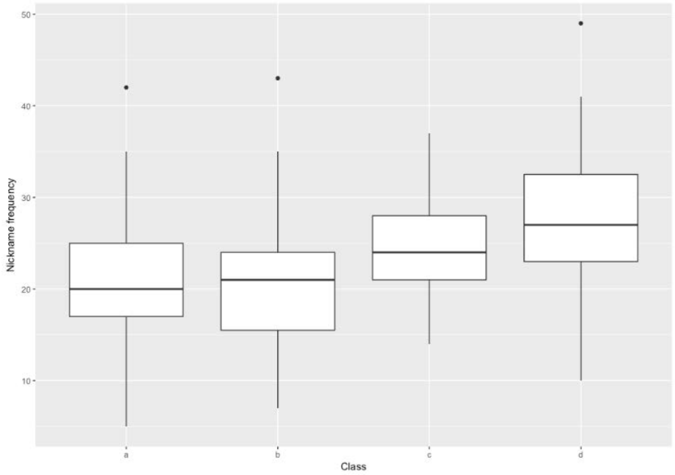
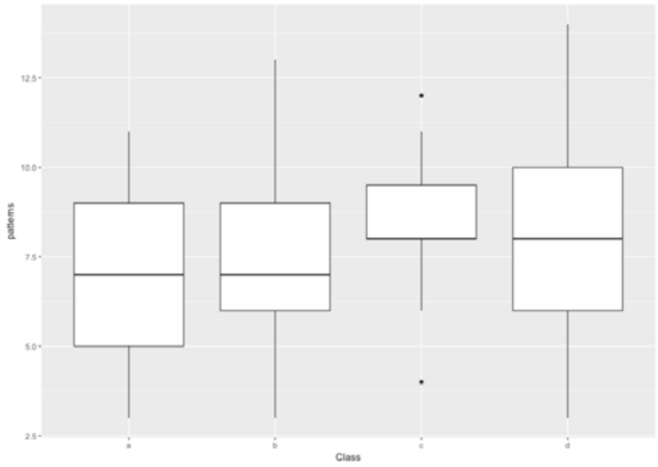


Figure 3:

Comparison of type frequency results taken from 4 different surveys.



After an Analysis of Variance test followed by a Tukey test to show similarity and differences, the results can be summarised as follows: Survey (d) is significantly different from (a) and (b) in terms of the number of tokens given, but the others are not statistically different as shown in the highlighted p-values given in the rightmost column of Table 3. In other words, paper (d) showed more student contribution by having them give a lot of nicknames.

Table 3:

Variance test showing differences of contribution with token frequency.

	diff	lwr	upr	**p adj**
b-a	-0.4285714	-5.0042004	4.147058	**0.9948948**
c-a	3.8857143	-0.6899147	8.461343	**0.1259916**
d-a	6.8285714	2.2529425	11.404200	**0.0009175**
c-b	4.3142857	-0.2613432	8.889915	**0.0722692**
d-b	7.2571429	2.6815139	11.832772	**0.0003700**
d-c	2.9428571	-1.6327718	7.518486	**0.3419914**

As for the type contribution, or the nickname pattern contribution illustrated in Figure 3. An Analysis of Variance test followed by a Tukey test showed different results. Sheet (a) and (c) are significantly different but the others are the same as seen in the p-value given in the rightmost column of Table 4.

Table 4:

Variance test showing differences of contribution with type frequency.

	diff	lwr	upr	**p adj**
b-a	0.6571429	-0.6744701	1.9887558	**0.5749169**
c-a	1.7142857	0.3826728	3.0458986	**0.0057324**
d-a	1.1142857	-0.2173272	2.4458986	**0.1350055**
c-b	1.0571432	-0.7682934	2.3887558	**0.1699306**
d-b	0.4571429	-0.8744701	1.7887557	**0.8085214**
d-c	-0.6000000	-1.9316129	0.7316129	**0.6455126**

Final Attempt

As the analysis project grew bigger with more areas and complications to investigate, the data was becoming more complicated to interpret and analyse. In some instances, there were many nicknames and no hypocoristics. In other instances, there were few names that showed patterns but cannot be grouped due them be a small set. For that reason, in the final approach the researcher did purposely point the students to the direction of the hypocoristic pattern that were questioned. This was done by giving an example in the beginning with only one name which had different patterns, *Basil*. The approach also was timed by giving the students 5 minutes to fill a sheet with 25 names.

Other Attempts

In between these attempts, other questions and directions of the research were encountered which required data collection. One of the areas, was a child data on hypocoristics. The child language investigation used a wug-like test where children were asked to provide with the nickname of 4-9 names. The children were first asked about their names and their siblings and parents. “What is your name? What is your father’s name, your sister, mother, brother?” Every time a name was given the interviewer gave the hypocoristic of that name and throughout the interview the hypocoristic of the child was used.

Once the children started to be less shy three different pictures were shown. The first picture was of a boy playing. The interviewer[15] told the kids that “this boy’s name is *Basil* and he is playing with his bunny; or he is playing on the slide. Do you love to play like him?” On the second picture the boy was with his mother and two of his siblings. The interviewer then starts by telling them who is in the picture. Then, informs them that “*Basil’s* mother calls him *Bassuul*. What do you think his mother calls his sister and brother?”. The interviewer then shows three other pictures with different people the only difference is that a nickname is not provided for the main person but asked about. In total, there were 4-9 names that were investigated depending on the child’s concentration and shyness.

[15] The researcher was accompanying a female interviewer (wife) who conducted most of the questions while video recording them using an Olympus LS-20m voice and video recorder.

The results which will be mentioned briefly in this thesis shows that children from age 6 can use hypocoristics similar to adults. However, it is not the one that they produce most. From 4 years old, their preference is to use the suffix –o after the name, as described below. This is in line with similar language acquisition research where children's first morphological production is witnessed in the addition of diminutive suffixes (Dressler, 2007). A summary of the collection attempts is given in Table 5.

Table 5:

Description of data collection attempts and their objectives.

Attempt	Names	Subjects	Objective
1	90	No count	Finding an approach and testing D&Z's claim.
2	45	30	Checking acceptability of some hypocoristics.
3	155	180	Collecting the main data for analysis.
4	25	67	Using data collection approach that targeted patterns.
5	4-9	93	Checking Children's awareness of hypocoristic formation.

Problems

There were few minor problems that occurred in all of the attempts. They are considered minor since they do not affect the general goal of collecting a variety of available nicknames. For example, in some instances subjects did not fill in the papers or they were not serious, so they just provided random nicknames or a funny name. In some instances, they did not have time to fill all of the paper. It was also pointed later that in certain classes non-Kuwaitis filled the papers. These issues are minimal for the rule-based analysis since the concentration of the data analysis focuses on types and not tokens. In other words, the total number of nicknames repeated and unrepeated, is not important but how many patterns is. On the other hand, it does affect an analogy-analysis as it is sensitive to tokens as will be explained. However, by having the different collection approaches in addition to the flexibility in the creation of the computational analogy, this becomes a minor complication.

There were other problems that did have an effect on the collected data. First, in every attempt the majority of subjects were urban Kuwaitis. Even those with a Bedouin family name that filled the questionnaires are considered urban speakers. This is due to having the collection be done at the English Literature Department and English Education Department where Bedouins are the minority. As a result, the data is representative of Urban Kuwaiti Speakers.

The second problem is also faced in many research attempts which try to capture an Arabic dialect. Kuwaiti Arabic is not a written language[16]. This resulted in complication with interpreting back and front glides. Back glides are written as [و] and they can be pronounced as either /wu/ or /uu/. This is seen with the name *fadwa*. When written in Arabic, the hypocoristic [فدو] is interpreted as either *fad-o* or *fadw-o* with a glide. The same thing is found with front glides [ي] where they can be pronounced two ways. For example, the name *hind* was given the hypocoristic [هنيد] this can be *hnayid* or *hneed*. This problem wouldn't occur if Arabic diacritics were used but it is usually the case that diacritics are not used especially with an informal form of Arabic.

The third and fourth problem are related to each other and has been discussed and studied as a problem for Arabic hypocoristics. Hypocoristics' acceptability is gradient as explained above with Prunet's finding. Due to that, in some classes where the surveys were distributed loud students were joking around and giving weird hypocoristics. This lead to other accepting it and writing it down. This occurred in the first attempt and the third attempt. In addition to that the unacceptability of a nickname is hard to point out even with a judgment task. However, there is one area where unacceptability of an item can be determined. This occurs only when a hypocoristic formation disobeys a structural constraint such as obligatory contour principle. In the data, this occurred widely when a glide occurs with a homorganic

[16] Although there are corpora that captures Arabic dialects. Nicknames are problematic for a corpus investigation and this is why from the start a corpus was not used and instead fieldwork approaches were used.

vowel. This resulted in not having */wu/, */uuw/, */ji/, and */iij/ in the data. The constraint also aided in interpreting some of the data that were problematic because of the Arabic written glides.

2.3 Hypocoristic Categories

The data is described on the basis of how they will be categorized for both the analogical analysis and rule-based analysis. As will be discussed later, giving a description that utilizes any phonological categorical group, natural classes, goes against the purpose of having an analogical theory of analysis since they are statistically driven. Nonetheless, as a general description approach, consonants and vowels will be used along with a Greek alphabet indexation along with segment numbering[17]. The name *badir* with the hypocoristic *badduur* will be described as $C^1V_0C^2V_0C^3V_0C_0 \rightarrow C^1aC^2uuC^3$, where C^1 is the first consonant taken from the name, C^2 is the second, and C^3 is the third. The vowels in the name are not specified nor indexed in the description of the name unless they are used in the nickname. For that reason, V_0 and C_0 are used to mean any number of consecutive vowels or consonants.

As for the vowels in the hypocoristics, /a/ and /uu/, they are not taken from the name but are part of the hypocoristic which is why they are specified. The specification of a sound will only be done when it targets a specific pattern. For example, to describe the name *ʔaħmad* which has the hypocoristic *ħammuud* the glottal stop will be specified, resulting in $ʔaC^1C^2V_0C^3 \rightarrow C^1aC^2C^2uuC^3$, since the pattern applies only to such names with glottal stops. It is used also with glides and they are written as Y instead of G because G will be used later for the analogy approach to specify Gutturals[18].

[17] The description approach is identical to Plag (2003) and Schane (1973) where they gave a linear rule-based templatic description. It has not been used here to avoid terms such as template, rule, and linear, which are associated with Generative Phonology.

[18] The appendix (A.1) has a list of symbols that were used since both R and Perl do not support Unicode characters.

CaCCuuC

Table 6 lists the first hypocoristic category. It is the default category since it has the majority of the provided hypocoristics. Under this category of hypocoristic, 14 variants are listed in the appendix. CaCCuuC is considered the default category for three reasons. First the majority of the names take this hypocoristic pattern; high type frequency. Second, the majority of the data is comprised of these; high token frequency. Not only that, but D&Z's work on hypocoristics was used mainly to analyse this pattern. This will be used as the name for the category in the analogy approach. It is organized like this purposely to show that analogy models can be interpreted as rule-based models.

Table 6:

Number of different patterns that are used as the default hypocoristic.

#	Pattern			name	hypocoristic
1	C1V0C2V0C3V0aan	→	C1aC2C2uuC3	ħamdaan	ħammuud
2	ʔaC1C2V0C3→	→	$C^1aC^2C^2uuC^3$	ʔaħmad	ħammuud
3	$C^1V_0C^2V_0C^3V_0C^4$	→	$C^1aC^2C^3uuC^4$	marjam	marjjuum
4	$C^1V_0C^2V_0wV_0$	→	$C^1aC^2C^2uuj$	marwa	marruuj
5	$C^1V_0C^2V_0$	→	$C^1aC^2C^2uun$	faj	fajjuun
6	$C^1V_0C^2V_0$	→	$C^1aC^2C^2uuj$	qaada	qadduuj
7	$ʔC^1tV_0C^2V_0C^3$	→	$C^1aC^2C^2uuC^3$	ʔbtisaam	bassuum
9	$mV_0C^1V_0C^2V_0C^3$	→	$C^1aC^2C^2uuC^3$	msaaʕad	saʕʕuud
10	$C^1C_0C^2V_0YV_0C^3$	→	$C^1aC^2C^2uuC^3$	baʃaajir	baʃʃuur

Reduplicates and CVCV

Four reduplication patterns appeared in the data. Interestingly, after looking at the various hypocoristic from different languages referenced here, one can conclude that reduplication is a cross-linguistic pattern of hypocoristic formations. The first one is a CVCV pattern. It appeared mostly in 5 variants. The most productive is C^1uC^1u. It appeared with almost every name.

There were a few that did not have this pattern due to a phonological constraint against gutturals, back glides, and glottals. The one case where they were not avoided was with *sˤusˤu*.

One reason could be that it is already a shortened name for a chick. Thus, their usage of it was not a use of hypocoristic but a use of a nickname which happens to be similar to a productive pattern.

Table 7 shows the different variants that appeared. The difference between how it will be described here and how the default hypocoristic was described is that it only applies to a name's first and second consonant. Thus, only the hypocoristic shape with the number of consonant will be used. In the fifth row, *mimo* appeared with names that have an m in any position. This also occurred with *fifi*, *zuzu*, and *ʃuʃu*. However, they won't be given a designated pattern category since they will be redundant information in most cases where these specified sounds are in the first or second position. Where they are not redundant is with few names; *fifi* for the name *hajfaaʔ*, *zuzu* for the name *marzuug*, and *ʃuʃu* for the name *ʕajʃa* where it is using the third consonant[19].

Table 7:

CVCV hypcoristic pattern.

#	Pattern	name	hypocoristic
1	C^1uC^1u	badir	bubu
2	C^2uC^2u	ħanaan	nunu
3	C^1iC^1i	fadwa	fifi
4	C^2iC^2i	ʔiimaan	mimi
5	mimo	ʔiimaan	mimo

The CVCV was not the only reduplicate hypocoristic pattern. Three other unproductive reduplicate patterns were also used and given in

Table 8. They are unproductive as seen in the data where they have a low token and type frequency[20].

[19] There were other CVCV patterns but appeared scarcely and thus will not be used.

[20] I believe that they are productive as a Kuwaiti native speaker I can easily form names with these patterns which all sound acceptable to me.

Table 8:

Reduplicate Patterns.

#	Pattern	name	hypocoristic
1	$C^1aC^2aC^3.C^2aC^3$	badir	badar.dar
2	$C^1VC^2.C^1VC^2$	badir	bəd.bəd
3	$C^2iC^3.C^2iC^3$	basma	sim.sim

Dimunitive ti/i

As mentioned above, that considering a pattern as a new type is based on formal and functional distinction. One area where this is useful is with the ti/i suffix. Prunet and Idrissi (2014) labelled them "Hypocoristic suffixes sounding like possessive suffixes" (p. 179). These affixal hypocoristics can be summarised as follows:

- They appear most of the time after a CaCCuuC hypocoristic (*bader → badduur-i*).
- -ti appears after female hypocoristics while –i appears after male hypocoristics (*baduur-ti*$_{f.}$ *and baduur-i*$_{m}$.).
- Most of the time they express diminutive when they are added to a hypocoristic.
- In few cases possessive is expressed when appearing after the name (ʕaziz → ʕaziz-i).

In their paper, Prunet and Idrissi (2014) clearly express that the behaviour exhibited with this hypocoristic reflects the complications found with formal linguistic analysis which do not capture functional features in their rules. They describe it as a "tug of war between formal analysis, where only phonology matters, and statistical distribution, where gender matters in [the choice of hypocoristic suffix]" (p.180). This suffix will not be further analysed as much more data is required to model which is an area where Prunet and Idrissi (2014) also see as a research area that requires more investigation.

-o Affix

While the CaCuuC formation is set as the default hypocoristic, the –o affix is the default affixal hypocoristic. Every name can use it. It is added after every name with two hiatus resolution strategies. The first strategy prevents the occurrence of a high vowel with -o by inserting a homorganic front glide in between [j] → [i#_o] The second strategy prevents consecutive back vowels by deleting the stem vowel [ɑ] → [∅] /#_o as given in Table 9 . This suffix is acquired before the rest of the hypocoristic pattern as noticed when collected from children. One point to note here is that the descriptive rules here are used as the rules in the rule-based theory given in the next chapter. However, they will be tested in the analogy approach as three affix patterns A= [#o], B = [jo], and C = [*Vo] as given in Table 9.

Table 9:

Productive -o suffix.

	rule	analogy pattern	name	hypocoristic
1	X → Xo	[#o] A	ʔaħmid	ʔaħmido
2	∅→ j / i __ o	[jo] B	hadi	hadijo
3	a → ∅ / __ o	[*Vo] C	huda	hudo

-aawi

Opposite of the –o suffix is the –aawi suffix. It is an unproductive suffix appearing after the left most closed syllable. Thus, it is concatenated after a shortening process of the name. The suffix appeared in the data with only 16 names which are all given in Table 10. What is interesting is that it appeared only in the first and second collection methods. This might be because it was trending at that time.

Table 10:

Unproductive -aawi suffix.

#	name	hypocoristic	#	name	hypocoristic
1	fadwa	fadaawi	9	ðˤuħa	ðˤuħaawi
2	ʕali	ʕalaawi	10	marwa	maraawi
3	3aada	3aadaawi	11	nada	nadaawi
4	maj	majaawi	12	rana	ranaawi
5	ʕiisa	ʕiisaawi	13	ranja	ranaawi
6	jaħja	jaħaawi	14	salwa	salaawi
7	maha	mahaawi	15	huda	hudaawi
8	nuura	nuurawi	16	ʕafaaf	ʕafaawi

2.4 Nickname Categories

Three of the hypocoristic categories given above are the main data given by subjects and some of them will be the topic of discussion as they constitute the productive formations which are part of the grammar. However, there were other nicknames given that need to be described since there is a role that some play in the given analysis. They also need to be mentioned since few have described Arabic nicknames extensively as presented here (Prunet & Idrissi, 2014).

Teknonym

Kuwaiti Arabic is one of the many languages that employ teknonymy as a nickname practice. Teknonyms are nicknames formed by taking the name of the oldest male child of a person and adding *abu* which means ‘father of’ in Arabic. What is interesting is that it can also be used even if a person is not a father. These cases occur with a few names where the teknoym and the name association are famous in a culture due to it being known from encyclopaedic information. For example, the name *ʕalij* has the teknonym *abu ħsein*. Everyone knows that because the Caliph ʕalij had a son called *ħsein*. These cases occur mostly with religious figures and historical ones such as kings and other leaders.

Semantic

A lot of research on nicknames divide them into external and internal where internal refers to hypocoristics because they use the name while external refer to the person. Hypocoristics refer to the name for phonological usage. Sometimes nicknames refer to the name for other associations. With teknonyms the association is with a religious figure that has that name making it an encyclopaedic association. With the CVCV hypocoristic *sˤusˤu*, a referral might be used to the word meaning of the noun *sˤusˤu* which is chick and has resemblance with being small as a meaning. Thus, even though it violates a constraint on gutturals it is still used. In addition to these associations there are associations which are phonetic but not considered hypocoristics. For example, the name *ʃeixa* rhymes with the hypocoristic *xoxa* which means ‘peach’ and it appeared 5 times in the data; the name *marzuug* was given *zigzag* which means the same thing in English.

On the other hand, what is referred to as external refers to any association with the holder of the name. Again, teknonyms can be used as an example. Teknonyms that are based on knowing the eldest child of the person are external. The reason is that such information is not found in either the name or from society. It is information that is about the person. Other external nicknames are like English *blondie* for a blond person or *lefty* for a person that uses their left hands. In Arabic, these appeared mostly as teknonyms but without a name after *abu*. Instead, it was a description. For example, *abu xaʃim* which means ‘father of a nose’. Or *abu ilbol* meaning ‘father of pee’. What is interesting is that they are most of the time derogatory nicknames in Arabic.

Bedouin

The Bedouin pattern is one of the nickname patterns that might fall in the hypocoristic category. First, they do show resemblance with the name. Second, there do seem to be following the diminutive formation pattern found in Classical Arabic where a back glide is inserted after the first consonant. However, they appeared scarcely in the data and formulating an analysis will be hard to achieve especially since it is collected from urban speakers[21]. For that reason, they are considered here a nickname. The association of this pattern with Bedouins is something that will be shown in the final chapter.

Other Name

There were other nicknames which cannot be categorised. There are two reasons for their existence. One is that they are the result of misinterpreting their handwriting. Two, has to do with a social aspect of nicknames. As a category, terms of endearment which is a type of

[21] The author is also urban and cannot form such nicknames except some of the ones provided.

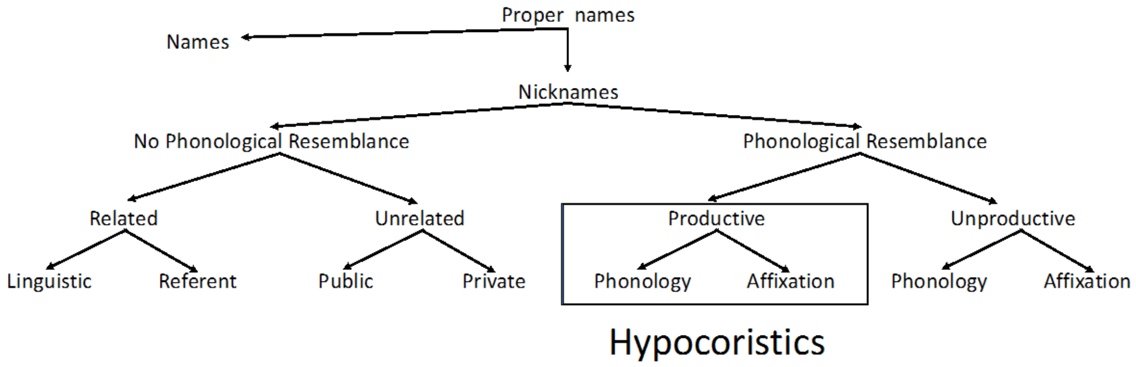
nicknames that are private names and not used in public. They are part of an intimate relationships that is not shared with others. Thus, this is another reason why some names appeared scarcely in the data and do not follow any pattern.

2.5 Naming Typology

Based on what has been presented here, a typology is proposed to cover the different types and to show which ones will be investigated in the following chapters. Having a typology for nicknames is not novel as it has been proposed before. The difference is when a linguistic typology is created it concentrates on the formal aspects of the nickname i.e. how they are formed (Taylor & Kennedy, 2015). On the other hand, in onomastic literature typologies focus on social factors of nicknames (Wong, 1997). The one in Figure 4 covers both and by doing so, the division drawn between onomastics and linguistics is not necessary. Instead, linguistic as a field would simply see what elements of names they want to study.

Figure 4:

Nickname Typology.



Proper names are the general category which has names and nicknames under it as a subcategory. Nicknames are what were described in this chapter. They can be divided into two categories. The first group are those that have no phonological resemblance with the name. Although they do not have phonological resemblance some have a relation with the name that can be drawn while others do not have any relation. The related nicknames can have a linguistic relation. For example, the teknonyms *jammaani* for the name *ʔajman* is considered to have a linguistic semantic relation based on how the name sounds. Although the relation can be seen as phonologically motivated, it is the semantics that the sounds trigger which is the basis of the relation. *ʔajman* sounds like the country Yemen and thus triggers the meaning of the country in its adjectival form, *jammaani*. Other examples given above, are *xoxa* and *zigzag* because they sound similar to the name. The other type of related nickname is when a relation is with the holder of the name as with the teknonyms that refer to the eldest child. It is based on referring to the referent as having that son. Table 11 has some of these nicknames from English.

Table 11:

Non phonological formed hypocoristics that refer to the referent

	Name	Nickname	Reason for reference
1	John	Reddy	John has red hair.
2	Sarah	Blondie	Sarah has blonde hair
3	Liza	Lefty	Liza used her left hand.

One category of non-phonological resemblance is left, and these are the random unrelated nicknames that one cannot draw any line between them and between the holder of the name. They might be known as with teknonyms that are part of society or a derogatory nickname. They might be the opposite as used in terms of endearment between couples or between secret societies such as Geishas in Japan. Table 12 lists some of the popular nicknames.

Table 12

Nicknames that are popular and used by anyone.

	Name	Nickname	Type of Nickname
1.	Any name	Polita	Popular Spanish nickname for any handsome person.
2.	Fatma	Batta	Popular Egyptian nickname for any female *fatma.*
3.	Maurice Richard	Rocket	Popular nickname given to a Hockey Player in the U.S.

On the other end of the typology are nicknames with phonological resemblance. They are not all hypocoristics. The reason for that is the productivity condition. In order for a nickname to be considered a hypocoristic, its pattern or formation needs to be shown on many names even, nonce names[22]. The unproductive ones will be considered in this typology as part of the small set that require further research to show whether they belong with the no-phonological resemblance or part of hypocoristics. In both the productive and non-productive ones, the resemblance can be seen and described as being a formation using affixation or phonological manipulation. One cross linguistic aspect mentioned in the literature to distinguish between the two, productive and non-productive, is their applicability to loan/foreign names (Katamba & Stonham, 2006). It is only hypocoristic formations that can apply across any given name. In this study, the analysis will be done only on the hypocoristics that are highlighted with the square; starting with a generative analysis.

[22] In the first collection attempt, a few nonce-names that were made up to resemble the structure of other names were included. They won't be discussed here as there was only a few.

3. A Generative Analysis

When using and defining Generative Grammar, a precision of the definition is required since the theory has seen various changes. For example, a great deal of criticism on the theory is discussed in sociolinguistic literature with regards to variation (Guy, 2011, 2014). Yet there are generativists who deal with variation even though Chomsky sees it as not part of grammar.

For that reason, this chapter will start with a brief history on Generative theory that lead to having two contrasted Generative rule-based approaches used in Semitic languages today. The two approaches have previously been used to analyse Arabic hypocoristics and they will be applied and evaluated using the data given in chapter two. A contrast between the generative approaches and the analogy analysis is spread over in chapter four and five.

3.1 Chomsky's Rules

Functional linguists and others who argued for analogy against a Chomskyan approach have always described it as a rule-based system. The term rules used in describing a language's grammar is not restricted to Chomsky's usage nor is it a feature that Generativists have inherited from Structuralists. It has been used to describe languages in every grammar since the Greeks. However, the way that Chomsky has placed it within Generative grammar is unique to language (F. J. Newmeyer, 1983).

In Chomsky's early proposal's the formation of complex words was restricted to syntactic and phonological rules. A syntax component was responsible for concatenating units stored in the lexicon and a phonology component was used to account for any allomorphic variation occurring to these units. For example, it was assumed that both regular and irregular past tense was formed by having both syntax and phonological rules apply. Thus, *played* and *wrote* were derived by a syntactic rule that would concatenate an abstract morpheme PAST with PLAY and WRITE. The PAST would then be realised as *-ed* or an *ablaut* to form the past tense by using phonological rules.

The assumption of the Standard Theory was that the syntax component of the grammar would produce rewrite rules similar to the ones given in Table 13. These included both the phrase structure rules (a-c) and the lexical item rules (d-f). As a result of using these rules the sentence in Figure 5 represented with a tree structure is produced. There is a final process where PAST and PLAY are handled by phonological rules that would spell out *played*.

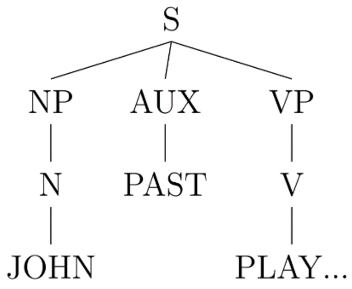
Table 13:

Rules in the Standard Theory.

a.	S	→	NP AUX VP
b.	NP	→	N
c.	VP	→	V
d.	AUX	→	PAST
e.	N	→	John
f.	V	→	PLAY; WRITE

Figure 5:

Tree Structure of Sentence formed by Standard Theory rules.



The example above, doesn’t show explicitness in Chomsky’s method which will not be attempted here[23]. The reason for not fleshing-out his work on word formation is because the approach set by Chomsky at that time “simply did not have adequate formal mechanisms for [...] [word-formation] phenomena” (Scalise & Guevara, 2005, p. 149). The “confront[ation of] the process of ‘internal computation’ [...] had Chomsky skirt[...] nervously” (Carstairs-

[23] The summary given here is written similarly in different papers and books (Anderson; Carstairs-McCarthy, 2002; Lieber, 2015; Scalise, 1986; Scalise & Guevara, 2005; Spencer, 1991). It is summarized here to show the concept of rules as an opposing theory to analogy. Although Standard Theory is not used any more, the way rules are set in it can clearly show a contrast to analogy.

McCarthy, 2002, p. 20). Whether Carstairs-McCarthy meant that Chomsky purposely avoided the issue due to its complications or avoided it by not concentrating on it, the theory left out how rules can account for word-formation.

It is important to remember that Chomsky’s theory was on syntactic formations and not words. It stressed the notion of regular productive rules which is the main contribution that Generative theories adopted, and analogical models rejected. Any grammatical sentence can be composed of a rule that combines an NP with a VP. On the other hand, words have more unpredictable behaviour that regular phrase structure rules cannot account for. This led Chomsky to separate syntax rules from word-formations and “call for a new, Generative, theory of morphology” (Spencer, 1991, p. 71). This theory of rules should operate in the lexicon prior to syntactic rules. Whether this requirement was important or not, “Generative grammarians were rather slow to respond” (Spencer, 1991, p. 73).

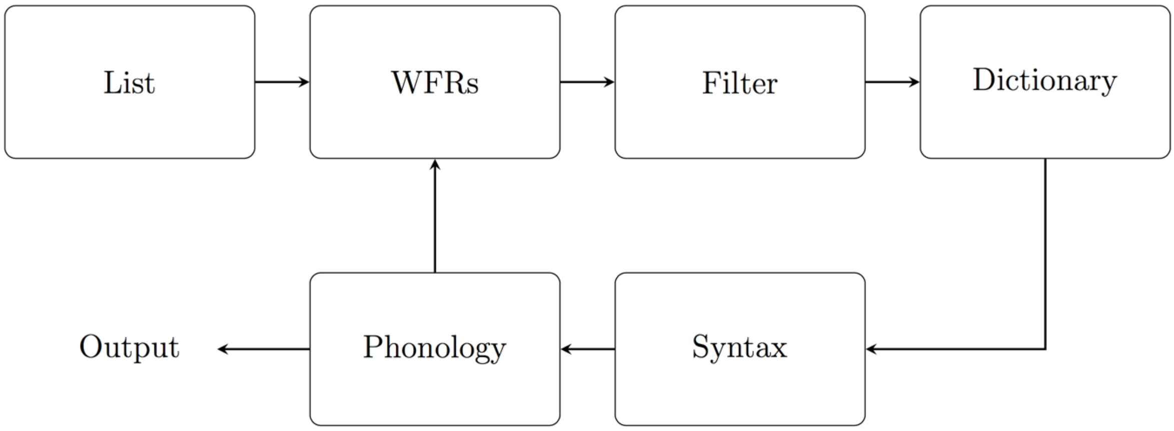
3.2 Halle’s Lexicon

Off the many generativists working at that time, Morris Halle took on the challenge with a rule-based lexical theory. Its main component was the lexicon. Instead of having syntactic rules responsible for word-formations as seen in Table 13, the lexicon acted as an autonomous component in the grammar where words are created. That way, the syntactic component was only responsible for phrase structure rules which inserted fully formed words in their specified positions.

Halle started with Chomsky’s mentalist approach to grammar. If a grammar mirrors a speaker’s knowledge about language, then it must show that a speaker knows (a) the meaning of words and (b) the existence of an ordered structure within a word. The latter knowledge entails that a speaker knows that a word such as *unstoppable* can be decomposed to *un+stop+able* and that it cannot be reordered to produce **un+able+stop.*

This is reflected in his grammar model as shown in Figure 6 where the four lexical components preceding syntax rules have a List component to store the non-decomposable units. These units are morphemes and are operated on using word-formation-rules (henceforth WFRs) located in another lexical component. The List complies with the meaning condition of speaker’s knowledge whereby each morpheme carries semantic details. As for the WFRs, they are responsible for native speaker’s other knowledge which is generating only well-ordered complex words.

Figure 6: Halle's Model.



Together these two components would generate a huge list of possible words in a language. For example, in the morpheme list a speaker would store *arrive, deny, derive, head,* and *-al*; along with their meaning; and certain idiosyncratic features that allow the WFR component to specify a set of morphemes over another that can be used as the base for rule application. Hence, a rule such as $[[X]_{v}.+ \text{al}]_{n}$. states that any unit in the list that is specified for being a verb can be combined with *-al* to form a noun. This would exclude the noun *head* which if the idiosyncratic details weren’t included would have produced **headal*.

Although, these two components do have the power to exclude many words that do not fit the structural description of the rule, there still will be some idiosyncratic details. For example, **derival* is a possible word that can be produced with the given rule. Yet, it is not acceptable. To solve the ‘possible but non-existent word situation’, Halle introduced another component that filters these idiosyncratic details. The Filter assigns features to any idiosyncratic element of words which would prevent them from being used.

Based on the processes that these three components achieved, Halle introduced a final component, Dictionary. The List of Morphemes, Word-Formation-Rules, and Filter, were all responsible for producing words that are listed in the Dictionary. It acts as a storage for the grammatical words in a language used by the syntactic rules to generate phrases. This final component completes the main requirement for Generative morphology that Chomsky called for, which is a lexicon component that would handle all of the morphological rules, leaving both syntax and phonology with universal context-free rules. What should be taken from both Halle and Chomsky’s model is summarised by Jackendoff (2002) :

> Since the number of possible utterances in a human language is unlimited, language users cannot store them all in their heads. Rather, [the][…] knowledge of language requires two components. One is a finite list of structural elements that are available to be combined. This list is traditionally called the “lexicon,” and its elements are called “lexical items”; for the moment let us suppose lexical items are words or morphemes. […] The other component is a finite set of combinatorial principles, or a grammar. To the extent that speakers of a language (or a dialect) are consistent with one another […] we can speak of the “grammar of a language” as a useful approximation to what all its speakers have in their heads (p.39).

3.3 Aronoff’s WB Formation

What Jackendoff stated above about having lexical items be for the moment as words or morphemes refers to one of the most debated topics in Generative Grammar. There was a consensus among Generative morphologists for a separation between syntax rules and WFRs, however there was a disagreement on what is the best unit for which WFRs can operate on. Halle’s model used morphemes as the base; resulting in Morpheme-Based (MB) lexicon. The opposite view expressed clearly by Aronoff took words as the base; resulting in Word-Based (WB) lexicon (Aronoff, 1976). The main argument of a WB approach came due to the complications that the irregular behaviour of morphemes creates for any WFR.

Up till now the term morpheme was used liberally to refer to any formative; whether it was an affix or a simple word. This is perhaps the closest application to how introductory textbooks define a morpheme: "Morphemes are the smallest individually meaningful elements in the utterances of a language." ((Hockett's definition as cited in Jensen (1990, p. 20)). Hocket's classical definition is a reflection of the Saussurean sign which Structuralists used as the base for their analysis and continued amongst Generativists (Aronoff & Fudeman, 2011; Carstairs-McCarthy, 2002; Lieber, 2014).

Structuralists' analysis was based on the decomposition of linguistic units, in search for **the smallest reoccurring formative that carries a distinct form and meaning**. For example, the words in Table 14 can be decomposed further based on the criteria of reoccurring distinct form and meaning. In (a) *phon* occurs in both words with the same meaning and shape. This reoccurrence of a single unit renders it a sign. On the other hand, *sand* has a similar form in both words given in (d), however, its meaning in Word 1 and 2 are different. In Word 1 it is 'a loose granular substance'. In Word 2 it does not carry any meaning by its on. Thus, it is not entitled a morphemic status. Instead, the plural form -*s*, *sand*, and *sandwich* are. This method of having a morpheme as a sign created the base which Halle used for the WFRs. They were stored in the List.

Table 14:

English Decomposition.

#	Word 1	Word 2	Decomposition
a.	phonology	telephone	phon, tele, logy
b.	slowly	badly	slow, bad, -ly
c.	badly	redo	start, do, re-
d,	sand	sandwiches	sand, sandwich, s

Abiding by the method given above in equating signs to morphemes for WFR, Aronoff showed that morphemes have an inconsistent behaviour that would lead to complications for WFRs that use morphemes as the base[24]. One example, comes from the cranmorph. In Table 15, the words in the first row show a tendency to be decomposed as *black*, *blue, cran*, and *berry*. However, if this is applied then *cran* would be left without a reoccurring meaning. While *berry, black,* and *blue* have a distinct **reoccurring formative that carries a distinct form and meaning** **cran* will be violating the condition of **reoccurring formative that carries a distinct form and meaning**. Although this issue was addressed by morphologists who adopted the morpheme as a sign approach, Aronoff rejected such an analysis.

Table 15:

Cranberry Morph.

#	Word 1	Word 2	Word 3	Invalid Decomposition
a	blackberry	blueberry	cranberry	black, blue, berry, *cran
b	refer	defer	transfer	re, fer, de, trans
c	remit	demit	transmit	re, mit, de, trans

Instead, Aronoff selected the word as the base which WFRs operate on. Neither do they have inconsistencies in representation of form and meaning, nor do they show idiosyncratic behaviour that morphemes show. As for the WFRs, Aronoff retained Chomsky’s rules that were used in Generative phonology and syntax. They behaved similar to rules in the Standard Theory. They have the power of copying, deleting, and inserting elements but with different restrictions.

Aronoff’s Generative Morphology used rules similar to the rewrite rules found in Generative Phonology (Chomsky & Halle, 1968). Since the rule used a word as a base to form a complex form, its input was always a variable acting as a placeholder for a set of simple

[24] The question of what constitutes a sign is a research topic formed independently from any morphological issues and even in this research topic some adopt a morpheme-as-a-sign and some adopt a word-as-a-sign.

words specified by formal linguistic features ($[X]_{\text{features}}$). The output was always a complex word that consisted of the position of the placeholder and the structural change ($[[X]Y]_{\text{features}}$). An example of three rules, are shown in Table 16

Table 16:

Word-Formation Rules.

#	Rule			Application			Type of Process
a.	$[X]_{\text{s. n.}}$	→	$[[X]X]_{\text{pl. n.}}$	kurdu	→	kurdukurdu	Reduplication
b.	$[[X]Y]_{\text{n.}}$	→	$[X]_{\text{n.}}$	shorudaa bakku	→	shorudaa	Truncation
c.	$[X]_{\text{adj.}}$	→	$[[X]\text{ness}]_{\text{n.}}$	happy	→	happiness	Insertion

Rule (a) is a reduplication process. It states that a word X which should be a singular noun would be recopied to form its plural. This rule is applied in Warlpiri to produce the plural form of *kurdu* ‘child’. Rule (b) is a truncation process. In Japanese, this rule is known as back truncation because it leaves the first nominal word of a compound word while retaining the composed meaning. Thus, *shorudaa* ‘shoulder’ plus *bakku* ‘bag’ will have the same compositional meaning, ‘shoulder bag’ even if *bakku* is deleted. Finally, rule (c) which is the quintessential rule given by Aronoff is an insertion process[25].

From the name of the rule processes one might conclude that an obscurity occurs between phonological rules and morphological rules. This is true and can be clarified. First, an insertion rule in phonology such as the insertion of a stop to break the nasal fricative sequence (as in *hamster* → *hampster*) is done not for lexical reasons. Instead it is an articulatory reason. Languages do not prefer such a cluster and instead breaks it up with another sound (Nathan, 2008). On the other hand, an insertion WFR is a lexical operation. It is applied based on the criteria of lexical formation; forming a new word. In short, both the phonological rule and the morphological rule parallel each other. The phonology rule provided by the phonology

[25] Stating that it is the typical rule is becasue Aronoff divided rules into different types. These differences do not affect the theory since they are still WFRs.

component inserts a specified segment based on phonetic and phonological features found in a language, while the morphological rule provided by the WFR sub-component is applied to form a new word.

This rule can also eliminate another obscurity. One might interpret rule (c) as a concatenative process of two lexical items, *-ness* and *happy*. This was the case in both Chomsky and Halle's model. For Aronoff, the *-ness* rule is stored in the WFRs along with the semantics, phonology, and the conditions required for its application. Hence, it maintains the classical definition of a morpheme but strays away from being one stored in the List.

In short, the status of generative morphology that is given here and has been taken as the instantiation mostly criticised by Skousen in his analogy approach, is that a grammar is a combinatorial system of placeholders which take **reoccurring formatives that carries a distinct form and meaning** from the lexicon to form new words. For example, with the hypocoristic *Tim* for the name *Timothy* a WFR would be truncation rule, $[[X^{\sigma1}]Y]_{\text{n.name}} \rightarrow [X^{\sigma1}]_{\text{n.hypocoristic}}$ where $[[X^{\sigma1}]Y]_{\text{n.name}}$ would be any name and $[X^{\sigma1}]$ would the left most closed syllable. If such a rule is applied to multiple names making a regular pattern and a productive rule, then it would be part of the grammar.

On the other hand, if it was applied to only one name or a small set then the extra-grammaticality argument will be used questioning the status of the rule and hypocoristics. For example, if the hypothetical rule $[[X^{\sigma1}]Y]_{\text{n.name}} \rightarrow [X^{\sigma1}]_{\text{n.hypocoristic}}$ was further specified to apply to one or a small set of names, as in $[[[X^{\sigma1}]Y]\text{aan}]]_{\text{n.names wich are used with adults}} \rightarrow [X^{\sigma1}]_{\text{n.hypocoristic,}}$ then such a rule will be considered not part of grammar.

This is why Arabic hypocoristics are a possible field for using the argument. As given, in the previous chapter, 16 different shapes exist for the formation of a single hypocoristic pattern that has the same function, which is what D&Z contributed to. They made the 16 rules into 1 context free rule abiding by the grammaticality condition of generativists, as will be shown below after giving a brief background on the tools they used- roots and templates.

3.4 Semitic Roots and Template

Along with Chomsky's rule-based formulation, Halle and Aronoff's lexical approaches shaped Generative morphology. The difference in Halle and Aronoff's proposal created a dichotomy in Morphological Theories, a MB and a WB analysis. The dichotomy played a role among Semitic Language analysis (Bat-El, 2011; Danks, 2011; Owens & Ratcliffe; Shimron, 2003; Ussishkin, 2011). Some Semiticists used WB, some used MB.

Anti-Consonantal Root and Template

WB approaches used Aronoff's condition on always having a word as the base for the word-formation process to analyse Semitic languages. A WFR would apply to a word creating another word. For example, causatives in Arabic are formed by a gemination process applying to the word. Table 17 lists examples of infinitive verbs that undergo a rule geminating the medial consonant to derive its causative[26]. This would entail that the infinitives are listed in the lexicon and a gemination WFR would apply to them. Another example given in Table 17 shows a nominalisation process. By using a melodic overwriting rule; by applying to the infinitive the nominal would be produced. The overwriting rule would lengthen the first vowel and raise the second vowel of the infinitive.

Table 17:

Arabic Word-Based Derivations.

#	Infinitive	Causative	Nominals	Gloss
a	katab	kattab	kaatib	write
b	ʤalas	ʤallas	ʤallis	sit
c	samaʕ	sammaʕ	saamiʕ	listen

[26] Arabic does not have real infinitives. It is usually the case that past tense stems are treated as infinitives. In Generative grammar posing an entity that never surfaces doesn't create a problem.

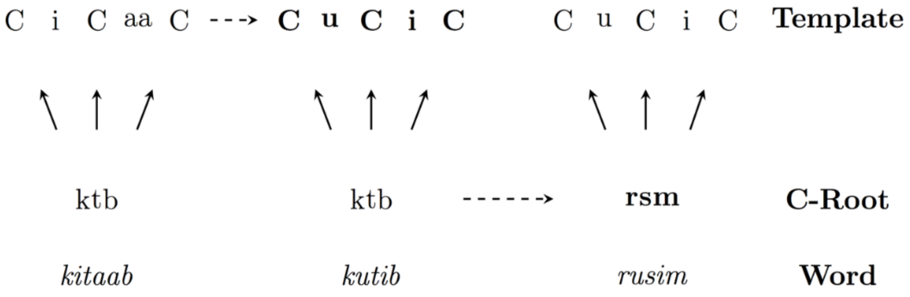
Pro Consonantal Root and Template

It has long been the case that Semitic languages had a non-concatenative structure. Words were taught as a combination of a root and a template. However, Generativists used a linear approach of analysis which prevented any possibility of re-establishing a root-template theory. This restriction ended with the introduction of auto-segmental phonology and prosodic morphology (Davis & Tsujimura, 2014).

Both advancements introduced Generativists to a possibility of restructuring words on multiple tiers as seen in Figure 7. In Semitic languages, words were thought of as a construct of two conflated tiers. The first tier is composed of a CV-template. The second tier contains a Consonantal-root. In isolation, they both carry a lexical meaning that satisfy Structuralists criteria for a **reoccurring formative that carries a distinct form and meaning.**

Figure 7:

Semitic Morphology.



As a result, Generativists gave each of these tiers a morphemic status which entails having them stored in the List. This made word-formation a process of conflating two tiers. In Figure 7 the dashed arrows show the change that occurs. *kitaab* ‘book’ is changed to *kutib* ‘written’ and then *rusim* ‘drawn’ by replacing the Template then the C-root. It is important to keep in mind that although the figure shows derivation in terms of changing tiers, it is not the case. The three words are formed in a similar fashion to MB concatenative processes. One can view it as a vertical concatenation or better as a tier conflation (J. C. E. Watson, 2002).

Another example, is given in Table 18 with three verbs that share the same template paradigm but differ in the root. In (a) the root *k-t-b* with the meaning of ‘related to writing’ is conflated with the templates *CaCaC* and *CuCiC* resulting in the active-passive contrast. The same in (b) and (c) with a constant meaning for the root *q-t-l* and *ð-k-r* meaning ‘related to killing’ and ‘related to reminding’, respectively. Unlike cranmorphs which were problematic for MB theories, the **reoccurring formatives that carries a distinct form and meaning** are clearly shown; raising a strong argument for posing them as a Saussurean sign. Thus, both were treated as a morpheme in an MB approach.

Table 18:

Arabic Verbal Paradigm

#	Active	Passive	root
a	katab	kutib	k-t-b
b	qatal	qutil	q-t-l
c	ðakar	ðukir	ð-k-r

D&Z’s MB Analysis

Arabic Hypocoristics is a great example of the debate between WB and MB lexicon in Semitic Generative Grammar. D&Z published the first description of Arabic Hypocoristic formation (Davis & Zawaydeh, 1999). D&Z used the tier-segregation analysis to support the mental reality of a morphemic consonantal root. They used a rule-based framework where the C-root of the proper name is used and conflated with a hypocoristic template[27]. Although there have been different approaches for the mapping between the tiers, D&Z opted for a segmental association of the C-root to the template which starts from the left edge and moves toward the right. They formulated it using indexation of the consonants. Their formulation is given as the rule, $C^1V(V)C^2V(V)C^3 \rightarrow C^1aC^2C^2uuC^3$ which is shown in Table 19 by applying it to certain names and it is read as follows.

[27] It is important to note that D&Z’s main discussion was the morphemic status of the C-root. Thus, they passed over any theoretical discussion on the template.

> Match the root consonants of the full name to the consonantal slots in the template $C^1aC^2C^2uuC^3$, where C^1 is the first consonant of the full name, C^2 the second consonant of the full name, [...] and C^3 the third consonant of the full name" (Davis & Zawaydeh, 1999, p. 90).

Table 19:

Hypocoristics derived using root and pattern.

#	Name	Hypocoristic	Root
a.	badir	badduur	b-s-l
b.	saalim	salluum	s-l-m
c.	mħamad	ħammuud	ħ-m-d
d.	majθaaʔ	majjuuθ	m-j-θ
e.	ʔibtisaam	bassuum	b-s-m
f.	ʕadnaan	ʕadduun	ʕ-d-n

D&Z's rule is accompanied by a repair strategy that occurs when an illegal segment is created by the morphology[28]. Since the template has a long back vowel, the conflation of a C-root with a final back glide leads to an absolutely neutralised environment, /uuw/. The reason why such an environment never surfaces is because of a cross-linguistic constraint that is avoided due to lack of a clear perceptual break between the back glide [w] and homorganic [u]. Instead, a front glide is used to create the clear perceptual break. As a result, one can view the derivation process that D&Z presented as a three-step approach sketched in Table 20.

Table 20:

Morphological Deirvation of D&Z's Analysis

xaalid	msaaʕad	sanaaʔ	marwa	fadwa	ʔarwa	Input
x-l-d	s-ʕ-d	s-n-w	m-r-w	f-d-j	r-w-j	**Root Extraction**
xalluud	saʕʕuud	*sannuuw	*marruuw	fadduuj	rawwuuj	**Template Mapping**
-	-	sannuuj	marruuj	-	-	**/uuw/ → /uuj/**
xalluud	saʕʕuud	sannuuj	marruuj	fadduuj	rawwuuj	**Output**

[28] The repair strategy is not part of the morphology. It is provided from the phonology component.

Ratcliffe's WB Analysis

D&Z did not give a MB analysis without taking into consideration a WB account. They wrote their paper using a rhetorical approach that leads one to reject a MB analysis[29]. By utilising non-linear morphology tools, they showed that one cannot formulate a single procedure, which applies to the proper name as its base in order to derive the hypocoristic. As an example, the structure of the proper names in Table 21 are different but share the same hypocoristic.

Table 21:

Hypocoristics sharing the same C-root

#	Name	Hypocoristic	Root
a.	ʔaħmad	ħammuud	ħ-m-d
b.	maħmuud	ħammuud	ħ-m-d
c.	mħamad	ħammuud	ħ-m-d
d.	ħamdan	ħammuud	ħ-m-d

Any analysis would be required to show how certain consonants do not appear in the hypocoristic. This is hard because first, the position of the deleted elements varies, and second, there is no phonological relation between what is deleted. In the former case, (a), (b), and (c) have initial consonants that are not used while (d) has the final consonant being deleted. As for the phonological and phonetic similarity, any rule that would prevent the initial or final nasal from appearing in (c) and (d) is not valid because there are many hypocoristics that are derived by carrying over nasals which appear in every position. In addition to that, even if one can show a WB analysis there would surely be more rules than the approach D&Z presented which will bring back the extra-grammaticality argument of having multiple contextual rules.

[29] Instead of having rule 4 reference underlying root consonants, they used the surfacing consonants.

This still did not stop Ratcliffe from responding. Ratcliffe from the start point rejects having a represented C-root in the lexicon as the base for morphological analysis. Like Aronoff’s view of Halle’s cranmorph, Ratcliffe sees the root as a deficient item exhibiting unpredictable behaviour. In his attempt, he used a theory different from the hypothetical WB analysis given by D&Z and even other WB theories.

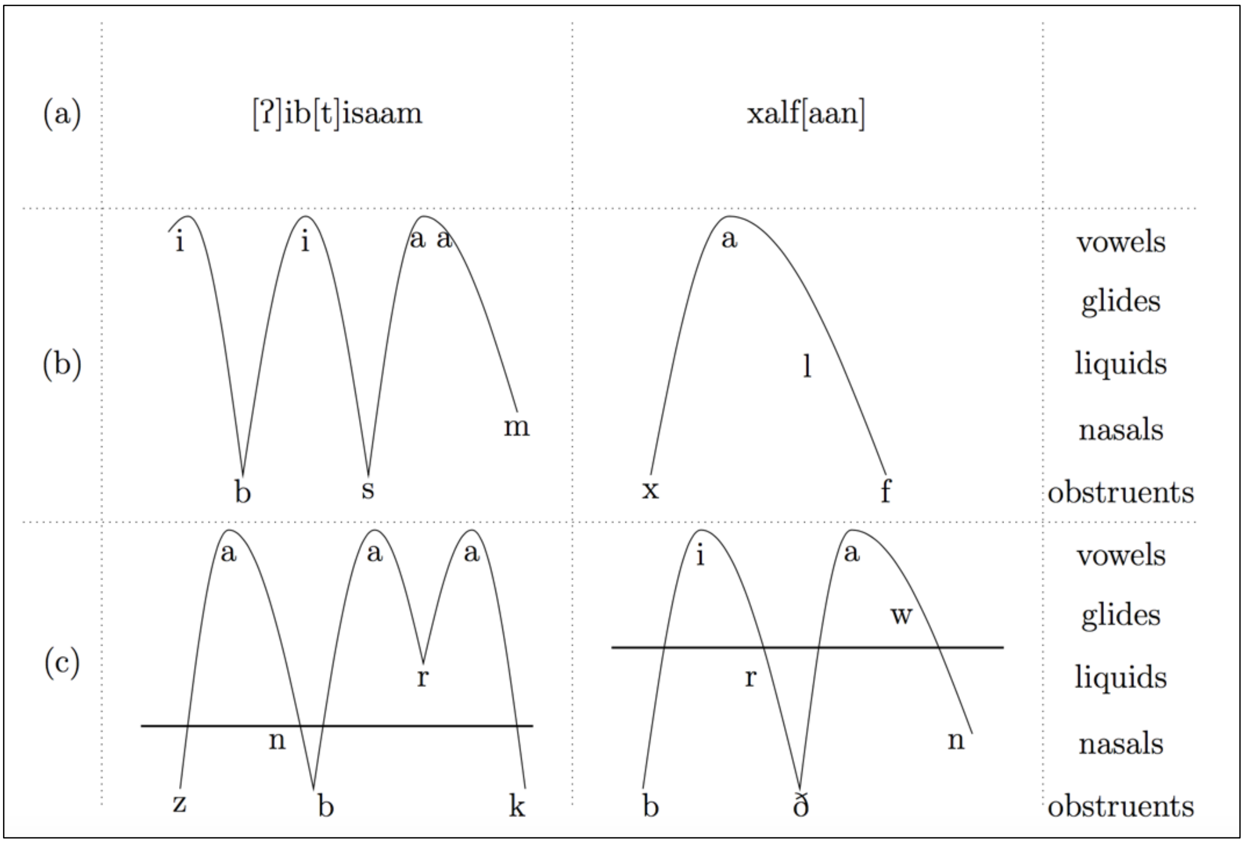
The rationality behind the theory is this: a prosodic morphological account is a process that operates on a word whereby material in the input of the derivation is carried over to an output and that material is defined phonologically using prosodic features such as moras, syllables, feet, prosodic words along with segmental features such as a CV-skeleton. Thus, why can’t one “expand[...] the notion “prosodically delimited” to “phonologically well- defined” and recognize[...] other parsing functions beyond prosodic circumscription, which identify or build up phonologically well-defined structures” (Ratcliffe, 2004, p. 78).

The hypothetical WB approach that D&Z presented was limited by the available tools used in Prosodic Morphology. By expanding these tools, Ratcliffe managed to show another phonologically defined structure that results from utilising the relative sonority of the segments. The process, which he terms Sonority Stripping, parses a word twice; first it parses for morphological structure if any and then it is followed by another parsing process based on a sonority hierarchy.

The result of the parsing is setting the segments of the word into two categories. Sounds that are not used for the derivational process and sounds that are used. He calls the latter a phonological root and maps it on to an “invariant shape, composed of slots for a fixed number of consonants, [...] that imposes or requires [...] [a specific set of segments from the word]” (Owens & Ratcliffe, p. 80). In Figure 8 an illustration of the process is given. Row (a) shows the first parsing process which locates affixes and epenthetic segments and prevents them from being mapped. They are given between square brackets. The glottal stop is an epenthetic consonant and the other two segments are affixes[30].

Figure 8:

Sonority Stripping



Once these segments are determined, sonority stripping is applied[31]. Ratcliffe formalises the process as follows: “from a given [...] word [...] parse out all syllable peaks, then all segments over sonority value S until a string of the necessary shape is obtained” (Ratcliffe, 2004, p.78). The first part of the formulation is represented by the wave in row (b) which shows the syllable peaks and troughs of the string left after the first parsing.

The second part of the formulation divides the segments over a variable sonority value until a string of the necessary shape is obtained. In row (c) the division is represented by the

[30] Ratcliffe defines an affix as “a string defined by constancy through a set of related words”(Ratcliffe, 2004, p.82). In other words, they are defined over a paradigmatic relation.

[31] It is not clear whether Ratcliffe treats the two parsing as one process called sonority stripping or just the second parsing process as sonority stripping.

horizontal line passing through the wave of two singular nouns[32]. Its location varies due to the template satisfaction condition. Since the plural form template, CaCaaCiC, requires four segments, then it is satisfied in the first word, *zanaabik*. However, with *birðawn* the bar needs to be raised[33] to obtain four low-sonority segments. Once the segments are determined, they are mapped to the template.

Row (c) also shows the main reason why Ratcliffe opted for a variable of sonority division rather than a categorical feature description. In the first name, the rhotic segment doesn't surface in the plural form *zanaabik* 'metal springs' but it does with the second word *baraaðin* 'working horses'. The reason is that template satisfaction requires three segments which leads to raising the sonority bar and accommodating the liquids. According to Ratcliffe, having a sonority-scalar division, rather than a theory of "discrete classes [allows] [...] Some segments [...] [to] belong potentially to both sets" (Ratcliffe, 2004, p.78). As a result, the answer to why certain high sonority segments surface in certain occasions and in other cases it doesn't, can be better explained using Sonority Parsing accompanied by template satisfaction.

The idiosyncratic behaviour of other high sonority segments, glides, is one area of research in Arabic that creates complexity in any analysis. Ratcliffe claims that sonority-based derivation is able to explain the issue better than MB theories. One such instance appeared in D&Z's data[34]. The names *ʕajda* and *dijma* have the exact same structure, CVjCa. However, it only surfaces with *ʕajda* resulting in *ʕajjuud* and *damduum*. Ratcliffe explains this problem by positing a processing constraint to coda glides with homorganic nucleus.

In Figure 9 the two words are parsed. With *ʕajda* the glide/nucleus distinction is clear-cut for processing. On the other hand, the glide in *dijma* is "invisible to sonority based parsing"

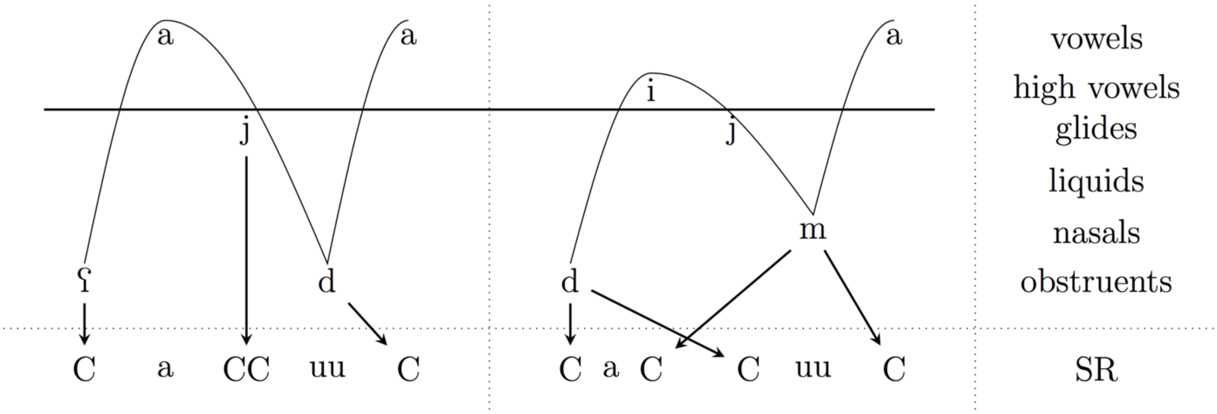
[32] The examples are taken from Ratcliffe for a better explanation since there aren't any similar structures with proper names.

[33] Ratcliffe explains the process the other way with the lowering of the bar. It has been given here the other way to show that the method Ratcliffe gives hasn't been fully sketched out. By having a template satisfaction condition demanding low sonority segments, there is no need for the bar. It is a redundant part of the grammar.

(Ratcliffe, 2004, p. 87). As a result, the mapping of only two segments instead of three can be justified by using Sonority-parsing. A MB analysis can’t explain this situation especially since the glide is in the same position of the C-root[35].

Figure 9:

Sonority Stripping of Glides



3.5 Evaluation of The Approaches

The two approaches given above have all been sketched out within a Generative framework[36]. To be more specific, the two approaches contrast on the exact formulation of the WFR and what units should be stored in the lexicon. However, they share Chomsky’s view of a modular rule-based grammar with context-free rules applying to the name. D&Z used morphemes as the base for the rule, while Ratcliffe used a word as the base for the rule. This is seen in the first section, left side of arrow, of the rule.

Since the two are all within a single theory, Generative Grammar, they share the same objective that can be used for an evaluation. As a grammar that mirrors a speaker’s competence they should be able to (a) generate only grammatical formations when applied and (b) account

[35] Note that in the Kuwaiti Arabic data it doesn’t need to be explained because *ʕajjud* and *dajjuum* were provided by speakers. Furthermore, the KA Arabic data that this is a mistake due to the multiple-choice questionnaire as *dajjuum* is the hypocoristic that a Jordanian colleague use for herself.

[36] Categorising Ratcliffe’s approach under a Generative theory can be debatable. However, he claims that sonority striping is a universal process that underlies morphology. In another publication, he places it within a rule-based grammar but links it to analogy-based grammars with minimal differences and a shared WB approach (Owens & Ratcliffe).

for every output produced by native speakers. These two conditions will be used in attempt to put the two approaches to the test. In other words, three questions need to be answered.

- Q1: How many hypocoristics are derived using D&Z's or Ratcliffe's approach are found in the data?
- Q2: How many hypocoristics are derived using D&Z's or Ratcliffe's approach are not found in the data?
- Q3: How many hypocoristics in the data cannot be explained using D&Z's or Ratcliffe's approach?

D&Z's MB Evaluation

Starting with Q1 for D&Z's approach; their rules when applied to the 165 names produce 140 hypocoristics that are all attested in the data and given in the appendix. Q2: Their rule overgenerates by producing hypocoristics that are not attested in the data as given in Table 22 and Table 24. The number of these are 28 and are given below along with possible explanation to why they do not appear. Q3: Here, it is the number of hypocoristics that subjects provided and a C-root and template analysis cannot explain. The number of these problematic are 36 and are given in Table 23. One reason given in the table for why some in the list do have a hypocoristic that D&Z cannot explain is from names that are borrowed. They are either Hebrew as in *juusif*, or Persian as in *ʔasiil*. Table 23 has data with high token frequency.

Table 22:

Names that are produced by D&Z's rule but not found in the data.

#	Name	hypocoristic	Possible explanation
1.	muntaha	*nahuuj	semantic blocking where nahuuj means the end
2.	ʔaħlaam	*ħalluum	semantic blocking where ħalluum is the name of a cheese
3.	btihaal	*bahhuul	semantic blocking where bahhuul is similar bahluul clown
4.	musˁtˁafa	*sˁaffuuj	gender blocking where sˁaffuuj is used with male sˁfaaʔ
5.	nɔf	*najjuf	gender blocking where najjuuf is used with male naajif
6.	foz	*fajjuuz	gender blocking where fajjuuz is used with male faajiz
7.	faj	*fajjuuj	

8.	jaħja	*ħajjuuj	There seems to be a constraint on /jjuuj/, /jjuur/, and /jjuul/ which share [+sonorant][+continuant].
9.	maj	*majjuuj	
10.	mniira	*najjur	
11.	ʔanwar	*najjur	
12.	nuura	*najjur	
13.	nawaal	*najjuul	
14.	nuurija	*najjuur	

Table 23:

Productive hypocoristic patterns that D&Z’s approach cannot explain.

#	Name	hypocoristic	possible explanation
1	ʔibraahiim	barhuum	no root
2	daana	dajjuun & dannuuj	no root
3	juusif	jassuuf	no root
4	leen	lajjuun & lannuuj	no root
5	marjam	marjjuum	no root
6	rula	ralluuj	no root
7	ʔasiil	ʔassuul	no root
8	ʔiimaan	mannuuj	
9	btihaal	battuul	
10	ʤawaahir	ʤahhuur	
11	ʤohara	ʤahhuur	
12	diima	dammuuj	
13	diina	dannuuj	
14	raaʔid	raʔʔuud	
15	zein	zannuuj	
16	saara	sarruun	
17	msaaʕad	msʕuud	
18	sultˤaan	saltˤuun	
19	ʔasmaaʔ	sammuuj	
20	sanaʔ	sannuuʔ	
21	miʃʕal	maʃʕuul	
22	ʃajmaʔ	ʃajmuuʔ	
23	sˤafaʔ	sˤaffuuʔ	
24	ðˤuħa	ðˤaħħuuj	
25	3aada	3adduuj	
26	mufiida	fadduuj	
27	maθaajil	majjuuθ	
28	ʔamaani	mannuuj	
29	maj	majjuun	
30	mniira	mannuur	

31	haadi	hajjuud
32	huda	hajjuud
33	miiʕaad	maʕʕuud
34	wafaʔ	waffuuʔ

Table 24:

Problematic names for D&Z that are produce with their rule but are not attested.

#	Name	Root	Predicted	Actual
1.	mniira	n-j-r	najjuur	mannuur
2.	nuura	n-j-r	najjuur	-
3.	nurija	n-j-r	najjuur	-
4.	ʔanwar	n-j-r	najjuur	-
5.	nawf	n-w-f	nawwuuf	najjuf
6.	fawz	f-w-z	fajjuuz	-
7.	haala	h-w-l	hawwuul	halluuj
8.	faj	f-j-ʔ	fajjuʔ	fajjun
9.	maj	m-j-j	majjuuj	majjuun
10.	muntaha	n-h-j	nahhuuj	mantuuj
11.	sultˤan	s-l-t	saluut	saltuun
12.	ʤawaahir	ʤ-w-h-r	ʤawhuur	ʤahhuur
13.	ʤawhara	ʤ-w-h-r	ʤawhuur	ʤahhuur
14.	nawaal	n-w-l	nawwuul	-

Ratcliffe’s WB Evaluation

Ratcliffe points to the borrowed name complication. It is one of the reasons why he disfavoured any extraction of underlying segments that are not available at the surface representation. He concentrated on showing this part from a theoretical perspective. As for its application, he resorted to showing only the “crucial data [...] [of] those cases where the consonantal string involved in a morphological operation cannot be [...] identified with the dictionary root” (Ratcliffe, 2004, p.77). The concentration on only showing the data that MB approach cannot account for or explain left some areas obscured for an evaluation. Ratcliffe doesn’t clearly give an answer to an important question which is how the first step of the analysis works.

In the first step a parsing process occurs with epenthetic and affixes. Ratcliffe defines an affix as “a string defined by constancy through a set of related words” (R. Ratcliffe, 2004, p.82). With hypocoristic data, all of the words are names and when put in a paradigm some names with final nasal will have the name be part of the hypocoristic (B) and others left out (A) as shown in Table 25. Thus, one cannot have a consistency of a string used for a definition of an affix.

Table 25:

Problematic Suffix parsing with Ratcliffe's approach.

	A			B	
	name	hypocoristic		name	hypocoristics
1.	ħamdaan	ħammuud	1.	ʔimaan	ʔimaan
2.	salmaan	saluum	2.	sultˤaan	sultˤaan
3.	ʕuθmaan	ʕuθθuum	3.	ħanaan	ħanaan
4.	xalˤfaan	xalˤfaan			

Ratcliffe admits to being selective with the data. It seems that his selectiveness isn't data-driven. In two cases Ratcliffe gives data-sets claimed to be problematic for a MB analysis. Yet, in the analysis he limits himself to just those items that adhere to his theory. For example, the two plural formation examples given earlier in Figure 8 was taken from McCarthy and Prince (1990, p.274). When going back to the original work one notices that there is another derivation, barnaamaʤ → baraamiʤ, that Ratcliffe leaves out which not only isn't compatible with his analysis but refutes it since it shows that the higher sonority rhotic was chosen over the lower nasal found in the input. He also does the same mistake with a data set that he compiled (R. R. Ratcliffe, 2003, p. 229). The data set is first given with several items. Then, in showing the application he lists just those that can be used to show the theory.

For that reason, one cannot evaluate it since it is not a full account. One reason for having an incomplete theory might be that he did not collect any data but resorted to using D&Z's set. Another reason is that his proposal should not be taken as a complete framework as he admits. He states that "the implications of a hypothesis which treats the root as a strictly-phonologically defined part of a word have been little considered [and he][...] wish[es] to consider only the starkest version of this hypothesis"(R. Ratcliffe, 2004, p.77). In considering all of the above, it is only fair to claim that he fell in his own criticism against D&Z in stating that their "papers [...] show [...] that there is a need for greater clarification and explicit formulation of the competing hypotheses if a meaningful empirical test is to be achieved." (Ratcliffe, 2004, p.73).

3.6 Conclusion

The chapter introduced A brief look at Generative Morphology. It discussed how Generative Morphology started with Chomsky's rule-based grammar and his call for a lexicon component listing all lexical idiosyncrasies. This led to two competing positions, MB and WB, which had as an objective the representation of the speaker's psychological reality. From these two main positions, various approaches and tools have been proposed. Some linguists took on the nature of the rule, while others took on the restrictions of the rules. The main thing is that a research field of Generative Morphology existed for linguistic insights from various languages. Within Semitic Linguistics, the main insight which was debated was the C-root and template. As a descriptive tool used by traditional grammarians, some Generativists sought to show its validity in a lexical theory of morphology while others took a stand against it. In the next chapter a novel approach will be presented to account for the hypocoristics. Furthermore, the CVCV pattern and the –o affix were not analysed here since they are repeating what has been said in the previous chapter. However, they will be analysed using the analogy approach.

4. An Analogy Analysis

Analogy as a linguistic derivational concept shares similarities with analogy as used in everyday language. However, it requires more restriction and limitation in order for it not to be very general nor opaque as defined by OED. For example, the second definition given in OED, ‘correspondence or adaptation of one thing to another’, leaves a generality of having anything be adapted from one to another without any restrictions. In this study, it is especially important since the word has been used frequently in linguistics and many criticism of these different analogy approaches are due to not having it be specific in its applicability (Itkonen, 2005; Skousen, 1989). To achieve a clear definition, analogy will be explained resulting in a restricted usage of the term that will be taken as the basis for the derivational analysis of hypocoristics which will follow.

4.1 Defining Analogy

The word analogy describes three intertwined things with a circular relation. It is used to describe a concept, a process occurring in the concept, and the result of the process. The three uses are all attested in the literature and some have given each distinct terms, based on the behaviour, static analogy versus dynamic analogy (Itkonen, 2005).

Analogy as a Concept

The concept of analogy is extensively used as a common day object and across several academic disciplines. It is used by parents to teach kids complicated issues which is best exemplified in the famous bee analogy. It is also used by great thinkers as an instructional tool for simplification. For instance, Einstein is quoted to explain the radio as follows “You see, wire telegraph is a kind of a very, very long cat. You pull his tail in New York and his head is meowing in Los Angeles. Do you understand this? And radio operates exactly the same way: you send signals here, they receive them there. The only difference is that there is no cat.”

("762: Analogies,") It ranges from a short sentence to a whole dialogue or story as with Plato's Meno and satires that appeared throughout history.

Whether it is used as a common day object or as a scientific tool it can be thought of as a single concept. Simply put, it is a concept that has at least two different groups containing different elements that are all related to each other for a specific purpose; resulting in having a parallel system. The analogy exists only when the knowledge of the relationships become known; this is stressed in Einstein's question "Do you understand this?".

This abstract concept of analogy has been given various representations using different paradigms, grids, and equations to assist in visualising the complex relation (Anttila, 2008; Blevins & Blevins, 2009; Itkonen, 2005; Skousen et al., 2002a). A widely used representation is the contiguity axis (also known as similarity axis, causal, or indexical) given below in Table 26 and Table 27. The tables show how two analogies can vary in size because of the number of elements that play a role.

Table 26:

Analogy of a two-place relation between a four-place relation.

	Birds	Fish	Functions
1	wings	fins	locomotion
2	lungs	gills	breathing
3	feathers	scales	protection
4.	beak	mouth	eating

In Table 26, an example of an analogy between two groups containing four elements is given. In the first group, Birds, there are four items that are related to Birds via a functional relation. In other words, wings and lungs are related to Birds because of a locomotion and breathing function, respectively. In the second group the same relation holds between Fish with fins and gills. The Fish column parallels the Bird column by having the same relations held between each item available.

In a way, there exists some sort of information that occurs between each element and the head of the Bird group which is transferred to the Fish group to create a parallel group. This is made clear in Table 27 where a blank column is presented. The range of items that can be added is wide enough to cover any moving object from ants and humans to trains and ships. For example, one can add a Ship group with sea and propeller in row 1 and 2 respectively. This will create a parallel group with the existing groups and result in an analogy of a four-place relation between a two-place relation. This aspect of having an undefined set is the most criticised feature of an analogy as it is the reason for not having a precise way of defining what goes there (Itkonen, 2005).

Table 27:

Analogy of a three-place relation between a two-place relation.

	Car	Airplane	Human	…	Functions
1	land	air	land	…	transport
2	wheels	wings	legs	…	move by

Going back to the definition of analogy given above, Table 26 shows ‘at least two different groups [Birds and Fish] containing different elements [wings, lungs, feathers, beak, fins, gills, scales, and mouth] that are all related to each other for a certain reason [locomotion, breathing, protection, and eating] that would result in having a parallel system’. The second part of the definition states that an ‘analogy exists only when the knowledge of the relationships become known’. This part is the one used to complete Table 27. Without knowing the function relation of each item within the group, adding any other group or item is impossible. This would result in having a group of random items listed in a grid rather than having an analogy.

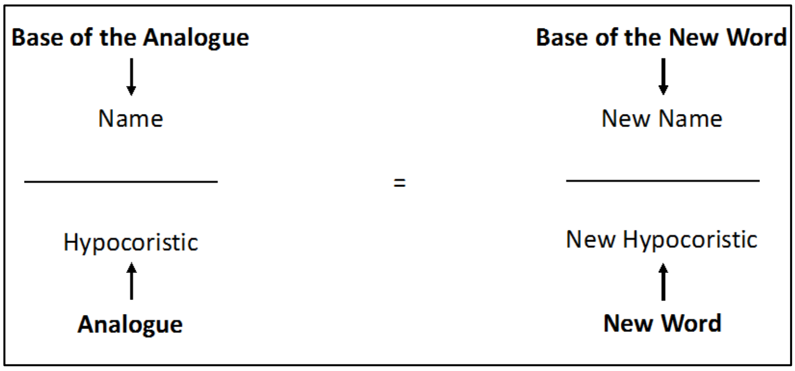
There are two more important features of an analogy that need to be addressed. First, it is the number of groups that matter and not the number of grouped items. A table with Birds and items along with their functions is not an analogy unless another group is added that mirrors the relationships found in the group. Second, although a wide range of items can be included not any item can be included. For example, one cannot add an immobile group to Table 27 because it cannot include any part of it that is responsible for the moving function. Thus, correspondences between groups is a necessity in an analogy.

Process of Analogy

In every analogy process, linguistic and non-linguistic, there are labels and terms that are used to specify the parts playing a role in the process[37]. There are four parts that play a role in a linguistic analogy as given in Figure 10. The ***analogue*** and the ***base of the analogue*** are the two items that form part of the speaker's knowledge. The speaker already encountered the two and stored them in an associative memory or lexicon. The third part is the ***base for the new word***. The base of the new word is either a new item not stored or a stored item that does not have any stored forms associated with it. In both instances, there will be a need to form the other part which is the ***new word***. The new word, is the result of the analogy process and the forth part of the equation.

Figure 10:

Parts of any Analogy Process.



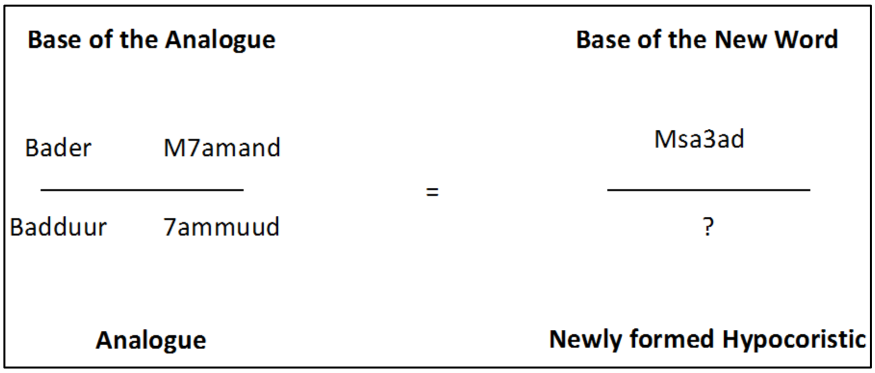
The psychological reality of analogy as exemplified above is a bit different. Hypocoristics are a great candidate to show the difference. For example, a speaker forming a

[37] Arndt-Lappe (2015) states that "there is no established terminology in the morphological literature to refer to most parts of the analogical equation" (p. 824). This is true and in previous writings, I used terminology that was adopted from Arab linguistic tradition which is based on Islamic theology literature. This has not been presented here for two reasons; recent linguistic publications that do have commonalities and Arndt-Lappe's terminology that do capture this commonality. Thus, a unified perspective is necessary for any further analogy research and it is taken here.

hypocoristic based on analogy will have as part of their memory that the hypocoristic for the name *bader* is *badduur* and for *mħammed* is *ħammuud* where the formation of the two are different. Then the speaker will encounter new names where the hypocoristics are not known. For example, the name *msaaʕad* will be the **base of the new word**. At a certain point, a process is required to form a hypocoristic for *msaaʕad*. However, there are two bases and analogues that can be used, and a question arise for which to use. The process that makes one choose a pattern over another can be a measured process using an algorithm. It is what analogy research and especially computational analogy is trying to provide and will be discussed below.

Figure 11:

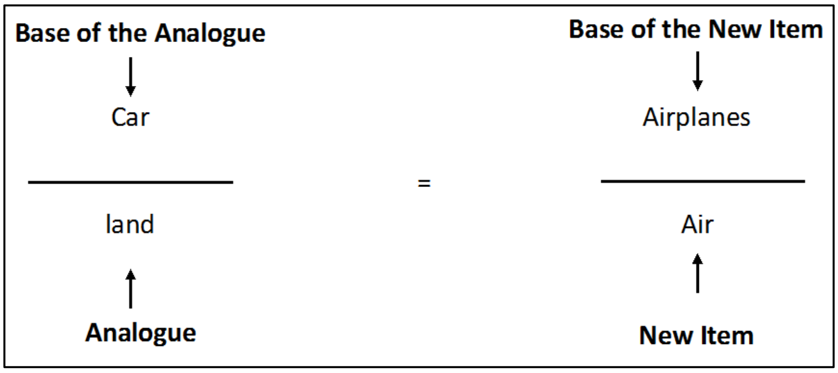
Example of Analogy Process with Hypocoristics



The terms can also be used with the examples given in Table 27 as illustrated in Figure 12. As stated above and rephrased here, the **base of the analogue** along with **analogue** have a relation defined on functional features. The **base of the analogue** also has a relation with the **base of the new item**. In knowing the basis for the relation between the **base of analogue** and the **analogue** one can use it to find the **new item**, air. In other words, by knowing that land is the means for cars to travel one can find the means for which planes travel and create a parallel system.

Figure 12:

Adaptation of Contiguity Axis to Analogy Diagram.



Referring back to the statement of analogy being an intertwined term, describing a concept, a process occurring in the concept, and the result of the process. Figure 11 show an example of analogy. The processes that is occurring in the figure is an analogy process. The result of the process represented as the newly-formed hypocoristic. Additionally, what has been illustrated has set out in the literature without a theoretical framework, but it is exactly how an exemplar lexicon in a usage-based theory works, which will be presented below.

4.2 Analogy in Linguistics

In the Preface of a bibliography on Analogy in Linguistics the authors observe that the history of analogy in linguistics resembles that of morphology as given in the previous chapter (Anttila & Brewer, 1977). It was one of the main tools used in language studies for a long time. It got moved to a peripheral status by the Chomskyan revolution. It then became a dominant tool that is used today in different linguistic analysis.

Existed

The linguistic tradition prior to Structuralism was mainly prescriptive. The objective was creating a grammar stemming from the description of a language with permission for the subjectivity of the grammarian. In other words, prescriptive grammarians prescribed and proscribed what they believed was the best language to speak even if it meant having rules adopted from foreign languages or prestigious dialects. This was done by having various linguistic tools used for forming the grammar and later teaching it.

Analogy was one of the tools that was used in both teaching and describing a language. The Greeks used it as an investigation tool in Sciences. Table 27 is an example showing how analogy has been used as a teaching tool and an investigatory tool where items that were left out of the axis were searched for by finding the relation that holds between the **base of the analogue** and **analogue**. Within linguistics, the Greeks used analogy in translations, phonetic descriptions, and even the development of an alphabet. An example of the three in one application is seen in the adoption of the Hebrew consonant aleph representing the glottal sound /ʔ/ to the Greek letter A representing the vowel /a/. The transference of information from one system to another clearly shows the utilisation of analogy as a scientific tool.

The analogy usage closer to the discussed topic, grammar, is seen with the Greeks'[38] insights on the nature and usage of the language. These speculations were part of *philosopha* which, unlike modern day philosophy, looked at every part of human knowledge. One of these areas that lead to a dichotomy originated from the search on the origin of language. On one side exists the school of thought that held regularity as a prevalent feature in the language due to analogy as a concept. Thus, linguists holding this view developed their linguistic descriptions and research using the process of analogy; this is seen in their development of paradigms that are still used today which describe language features such as tense, case, conjugation, derivation, etc. These paradigms share all of the parts found in an analogy and were used to complete each part. Greek Grammarians even reached a point where semantic

[38] A historical difficulty arises in connecting the lines between thoughts and thinkers. Thus, instead of having ideas linked to Aristotle or Plato the word Greeks would suffice even though most of the discussion is traced to the Stoics.

meaning was imposed on a word due to it fitting a formal paradigm based on structural relation. This was the start of morphological paradigms.

On the opposite end are those who viewed irregularity or anomaly as a prevalent feature of language that rejects the analogists' equation of one word, one meaning. Thus, their work in language resulted in showing how one word has multiple meaning depending on context (Robins, 1997). This is mentioned here because the anomalists tolerance for irregularities in a language has introduced to linguistics a field where the search of variation is crucial in the development of a grammar; the analogy framework given below also considers this issue and thus can be seen as middle platform in the dichotomy. Thus, it is not the opposing ends that are important in linguistics, but it is the questions that were investigated as a result of having these views. They marked the beginning of linguistics as an investigatory field as reflected in the nominated analogy framework.

The Greeks are in a way representative of Western Linguistics where other frameworks that appeared later were influenced in one way or another. As commented by Robins "the Greek thinkers [...] initiated in Europe the studies that we can call linguistic science in its widest sense [...] from, ancient Greece until the present day" (p.12). The existence of analogy-anomaly dichotomy is not considered valuable by some. However, when looked at as a theoretical platform from which other held views on language fall under it, then it is not to be taken trivially. This view should be held especially if the analogy that they used is seen as the same one which has continued to be used even amongst the many descriptive schools of linguistics that appeared from the Romans, to the neogrammarians, and up to the taxonomies of American Structuralists.

It is debatable whether Greeks use of analogy influenced Arabs especially in their work on logic, but analogy strongly existed as a tool in theology and linguistics[39]. In theology, it basically was a tool which was used to arrive at a verdict on whether something is forbidden. For example, beer wasn't known to Arabs. The verdict on having it be forbidden was drawn from an analogy with wine whereby both form a parallel group with liquid and drunkenness.

In linguistics analogy resulted in the emergence of two main schools in Arabic Grammar[40]. Qiyaas which is translated as measurement or metrics; refers to analogy which was a tool that the grammarians used. The two schools that emerged debated the validity of using analogy as a tool in grammar description. Kufis claimed that in cases where a linguistic entity is not known analogy is motivated to find it without any restrictions. This was opposed by, Al-Basris who wanted old texts, the Quran, Poetry, or Arab speakers from specific remote regions to be the decisive factors[41]. Again, it is important to note that this process is the exact same analogy process that will be used as the nominated research approach with hypocoristics; transferring what is known to the unknown.

Marginalized

Analogy was marginalized and the main reason behind that was Chomsky (1986) as he clearly states "that there is little hope in accounting for our [linguistic] knowledge in terms of such ideas as analogy"(p.12). His direct attack that pushed analogy away from linguistics comes directly from his poverty of stimulus argument. It is fair to say that his argument "still [holds]today [and] it provides the basic rationale for the entire generative enterprise" (Itkonen, 2005, p. 68).

[39] There is no doubt that Arab grammarians were influenced in many areas by Greeks in the era of translations. However, the role of analogy in grammar and science is presented by historians differently. One point is that it never existed until contact with Greeks. The other is that the concept of analogy was a theological tool used amongst Islamic scholars which later was adopted in other sciences.

[40] Many of the literature on linguistics mention two but theology references on analogy in language has a variation occurring between the two schools.

[41] Subjectively, it is interesting to mention that modern day literature on Arabic is beginning to show influences from Chomskyan Grammar as an opposition to traditional schools of language. In a way, historical developments repeat their directionality in different regions at different times.

The poverty of stimulus argument which is also known as Plato's problem states that the environment from which a child acquires the language is in no way representative of what the child produces. In other words, a child's output does not equal its input. Two things related to analogy stems directly from this argument. First, according to Chomsky this entails that any external effect of any sort does not and should not play any role in the grammar.

Second, a requirement for this problem emerges in Chomsky's innateness hypothesis. If mental grammars are not developed based on external languages, Chomsky's hypothesis is that it must be innate. Having language be innate and the proposal of an I-language description is the pivotal point in linguistic history. Any search external to language does not constitute part of grammar. What should be searched for is the initial state or what the rules that the child knows prior to any language encounter. As noted in the first chapter, this not only affected the way linguistic data is used and analysed, but it had an impact on the type of data that is relevant to linguistics.

Reintroduced

Anttila and Brewer (1977) comments that analogy "has received the renewed attention of linguists, including […][Generativists], in recent years, after having been dismissed by Chomsky and his followers as of little use in matters of linguistic theory" (p.VI). The renewal was brought forth by few linguists, cognitive scientists, and philosophers, that never joined the Chomskyan Revolution along with others who reacted against the revolution due to certain inconsistencies in the Theory (Geeraerts, 2006; Littlemore & Taylor, 2014).

4.3 Usage-Based Theory

The reintroduction of analogy can be seen in the various work that has recently been published (Arndt-Lappe, 2015; Blevins & Blevins, 2009; Itkonen, 2005; E. Mattiello, 2017; Skousen et al., 2002a). While Skousen concentrated on placing it as the sole tool for derivation, others place it along with categorisation, chunking, rich memory storage, and cross-modal association, as part of a cognitive toolbox that take part in the many aspects of human behaviour from vision to riding a bike (Bybee, 2010). Whether it is the approach taken here, Skousen's work, or the approach of others using analogy, analogy is an implementation belonging to Usage-based theory.

A Usage-based theory is based on having the use of language be at the core of any linguistic structure. In other words, a person's everyday linguistic experience is what creates a speaker's mental knowledge of language that is used to create morphological, syntactic, and phonological formations. A linguistic experience can best be understood as the sensitivity to contextual information accompanied by any statistical detail of linguistic forms. In other words, context and frequency.

A Usage-Based theory treats the lexicon and the formations as emergent which presupposes the existence of variation in the data and allows for a change (Bybee, 2010; O'Grady, 2008; Pierrehumbert, 2001; Rácz, Pierrehumbert, Hay, & Papp, 2015; Su, 2016). An emergent grammar is a changing entity that maintains a developing phase which originates from previous states and recent encounters. As an analogy of simplifying emergent grammar it is like a "new machine built out of old parts, [new parts, and obtained parts]" (O'Grady, 2008, p.449)[42]. Thus, the grammatical theory strives on explaining variation and irregularity in language which is a dominant feature of hypocoristics. They can be explained by context and frequency.

Exemplar Lexicon

A usage based-grammar takes it that language is exemplar-based. In reference to Generative grammar, the main point of difference is with storage and processing. Chomsky's approach in generative theory sees derivations as the result of highly advanced processing task.

[42] Although the analogy has been quoted a few times and referenced to Bates & MacWhinney, it is not found in their paper.

A child acquiring a language would do it by first going through a huge amount of data from the environment and then processes it to find various generalizations in that data. Then the child would use them to set the parameters and match them to various context free rules that are already stored in the mind as part of their biology. It is when the child produces the language that these new stored elements are used. As discussed in the previous chapter, the lexicon would have minimal number of idiosyncratic units that a rule requires to produce sentences. This whole process was an effort to minimize any pressure on the storage at the expense of having maximized processing.

From the start-point, in an exemplar theory, the lexicon is treated as a store house that encompasses detailed linguistic experiences that includes various units, morphemes, words, and phrases. The units can also be stored redundantly. The supporting evidence to having such a storage ability has been shown in various work including Generativists' who showed that the lexicon can store words and phrases (Jackendoff, 2002; Pinker, 1999). In fact, various psycholinguistic evidence suggest that words are not stored as morphemes but as whole units (Bybee, 2001, 2010; Eddington, 2009). For example, in a psycholinguistic study, subjects showed different reactions when they were tested with whole words but with different manipulated speaker-speech-related factors such as pitch and speed (Kolers & Roediger, 1984). The study suggests that it is the whole word that is stored and not morphemes since the difference in reaction can only be interpreted if the word was stored.

The claim that is made here with regards to Exemplar-based findings of the lexicon and hypocoristics is that a speaker would store every name encountered along with a nickname/hypocoristic as closely associated units. These units can have features defined over statistically driven elements (Saffran, 2003; Saffran, Aslin, & Newport, 1996). In other words, the sound /ʤ/ is not interpreted as two segments because statistically /ʤ/ appeared in more contrasting contexts where a difference can be noticed. This type of processing where items are stored and matched according to frequencies in the data is an analogy process and will be shown in the next section by using a computational approach.

4.4 Computational Analogy

The concept of analogy introduced today as part of exemplar theory has seen a huge support and experimentation from computer scientists especially with the access of large electronic corpora and super processors (Joan, 2007). Although the implementations and work has seen a huge development they can all be viewed as a form of what has come to be called k-nearest neighbour models.

Prediction Algorithms

A better way of understanding how k-nearest neighbour and prediction algorithm works is by looking at the field that they are most used in, machine learning and data mining (Witten, Frank, & Hall, 2011). One objective of the field is to find patterns from large body of any available data. However, the main goal is to create better prediction models. One model that is created in data-mining is a weather model to predict the weather. A simple model which has the sole purpose of predicting rain can be as simple as having a set of inputs associated with outputs listed in a data-set. The input could be a set of features such as cloudy, windy, and October which are associated with the output raining. The input and output are reminiscent of the **base of the analogue** and **analogue** which make up an exemplar.

Table 28:

Database of Model for Weather Predition

	Input			Output
	cloud status	**wind status**	**month**	**rain status**
1	cloudy	not windy	October	raining
2	clear	not windy	December	not raining
3	cloudy	windy	July	raining

Each row in the model is an example of one item that contains an input and output. The output can be thought of as a dependant variable that is the result of the input which is the independent variable. The input contains three variables, cloud status, wind status, and month of year. Each of these variables have a specific number of attributes. The cloud status has two attributes[43], cloudy or clear. The wind status also has two attributes, windy or not windy. The month has 12 attributes with the months of the year[44]. The output is also stored as a variable with two attributes, rainy and not rainy.

The model's purpose is to predict the rain status of any new day based on the attributes of that day. For example, on a clear and windy day in July, will it be raining or not raining as given in Table 29. An algorithm such as k-nearest neighbour, AM, or regression will be used to look at the data and classify it by finding the attributes of each row and then nominate one of the three given items/rows in Table 28 as being the closest to the new day and give it the same output of that row. For example, in the model given in Table 28, row 1 has zero shared attributes with the new day, row 2 has one shared attribute, and row 3 has two shared attributes making it the closest to the new day. For that reason, the new day is predicted to have the same output of row 3, raining. This measure is the basis of every different algorithm. There are always a number to weight or count in order to arrive at which item is the closest. This is even attested in Skousen's approach where the number of differences rather than similarities between items is the measured amount to determine the output.

Table 29:

Testing Instance of a day with an Unknown Output Status

	Input			Output
	cloud status	**wind status**	**month**	**rain status**
1	clear	windy	July	?

In reference to the analogy terms used above, after connecting the **base of the new word**, clear-windy-July, to the **base of the analogue** cloudy-windy-July, the **new word,** will

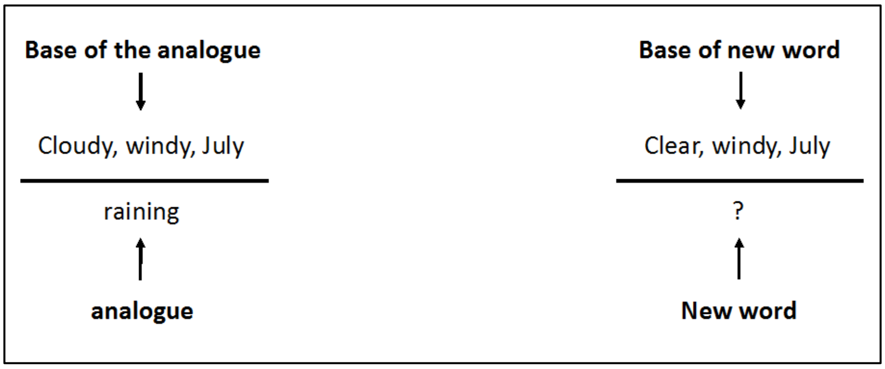
[43] Skousen calls them features.

[44] Actually there are 4, 2, and 13 where an additional null attribute is available. Thus, it is windy, or not windy, or not known.

follow the same **analogue** pattern, raining, that is associated with the **base of the analogue**. The **base of the analogue** and the **base of the new word** in the linguistic analogy examples were given as a single word. With the weather model, they are given as three attributes as shown in Figure 13. This is not so different from having a single word. The name Sam which is a single name can also be written as (consonant, vowel, consonant) or (CVC) or any other category that replaces the sounds in Sam (123).

Figure 13:

Weather model depcited in the linguistic analogy formation



Types of Predictions

In short, a computational model is used to predict an unknown output for a given input by using an algorithm which searches through related data. Once the predicted output is found, computational models have two ways of presenting the result. The first method is similar to what was shown in Table 28 and Table 29 with the weather model. One winner output is chosen, it is raining.

The other way that a predicted output is given is through probability percentages. It will instead be the percentage chance of having it occur or rain. For example, row 2 in Table 28 has one out of three attributes that are shared by the new day, this will be calculated as follows: ((1/3)*100=33) which means there is a 33% chance that it will rain; and row 3 has 2 out of 3 shared attributes ((2/3)*100=67)) which means that there is a 67% chance that it will rain, as given in Table 30.

Table 30:

Example of two types of results that a model can provide.

Model A	Winner	Model B	Probabilities
Result	Rainy	Result	33% will rain 67% will rain

Types of Model Evaluation

Now that a model is created and an item with an unknown output has been predicted using either a winner takes all result or a probability of occurrence, what is left is evaluating the model's success. How good is the model at predicting unknown outcomes? In other words, how good is the weather model in predicting rain? To do this there are two metrics commonly used to evaluate the accuracy of models. The first one is a simple accuracy metric. This metric is used with models like Model A where the results are given as a winner.

It simply compares the accuracy of the predictions to the actual data. So, for example, if a list was made of a whole week where a person would check predictions of two different models each day and lists them along with the actual weather, it will look like Table 31. The accuracy metric is simply the number one gets when the correct predictions are divided by the total number of items and multiplied by 100. Thus, Model P has 57% success rate. Model Q has a success rate of 71% making model Q a better model.

Table 31:

Comparison of the success rate of two different models.

Days	Prediction Model P	Prediction Model Q	Actual Weather
Saturday	Rainy	Rainy	Rainy
Sunday	Rainy	Not rainy	Not rainy
Monday	Rainy	Rainy	Rainy
Tuesday	Not rainy	Rainy	Rainy
Wednesday	Not rainy	Rainy	Not rainy
Thursday	Rainy	Not rainy	Rainy
Friday	Rainy	Not rainy	Not rainy
	4/7	**5/7**	
Success rate:	**57%**	**71%**	

As for models like the probability model given in Model B. These are compared with metrics used to compare two probability distributions. The two that will be mentioned in this study are cross entropy and Kullback-Leibler Divergence, often termed as KL divergence (Kurt, 2017). They are both the same, but KL is modified to have a true zero representing the actual data. In other words, in both metrics one can evaluate whether one model is better than the other by looking at the two numbers and the one closest to zero is a better model. However, there is no true zero where distance can be measured with the optimal prediction. The true zero has only been added to KL as a modification[45].

Table 32:

Step one in evaluating models using cross entropy.

	Subject Response (A)			Model Prediction (B)		
	C^1uC^1u	C^2uC^2u	C^1iC^1i	C^1uC^1u	C^2uC^2u	C^1iC^1i
besme	8	9	4	60%	20%	20%
sene	8	4	0	80%	13%	7%
hedpl	0	5	0	0	30%	70%

The weather example above is not a good dataset to show how the metric works. Instead, Table 32 shows an example for the CVCV hypocoristic pattern[46]. The table has data from subject responses that appeared in the questionnaire (A). This data contains the token frequency of each pattern given for that name. In other words, in the questionnaires there was a total of 8 *bubu* and 9 *susu* and 4 *sisi* for the name *besme*. On the right side of the actual data is what the model predicted. This was done by removing the three names from the model's data-base and then having them be in a test file to question their prediction. Thus, the model does not know what the output was for the three names.

[45] Since cross entropy is used more commonly with model evaluation and has many online tutorials it will be used here.

[46] The name in the table and in other places are sometimes written by not using IPA since it is not supported in Perl or R. The codes for Arabic are given in the appendix. Where data file is referred to the non IPA transcription will be used.

The model just processed a file with three names that do not have any output associated and then searched through the hypocoristic database which only has the CVCV hypocoristics and gave the prediction results in probabilities. Thus, for the name *besme,* the model predicts that in 60% of the times it will get the pattern C^1uC^1u. 20% of the time it will get C^2uC^2u and 20% of the time it will get C^1iC^1i.

The second step taken after getting the models prediction is to compute the probabilities. The purpose is to transform the data into a format where they can be comparable. The pattern C^1uC^1u for the name *besme* will be 38% (8/(8+9+4)), the pattern C^2uC^2u for the name *besme* will be 43% (9/(8+9+4)), and the pattern C^1iC^1i for the name *besme* will be 19% (4/(8+9+4)). This is done with all of the subject response. All of these results and the model's prediction (B) data will then be divided by 100 and will result with the data given in Table 33.

Table 33:

Step two in evaluating models using cross entropy.

	(A) Subject Response			(B) Model Prediction		
	C1uC1u	C2uC2u	C1iC1i	C1uC1u	C2uC2u	C1iC1i
besme	0.38	0.43	0.19	0.6	0.2	0.2
sena	0.67	0.33	0	0.8	0.13	0.07
hadpl	0	1	0	0	0.3	0.7

Finally, the average of the cross entropy is given. It is taken by calculating cross entropy which is what is found when comparing the two distribution probabilities between (A) Subject response and (B) Model prediction for each name. For *besme* this would mean comparing 0.38, 0.43, 0.19 with models predicted distribution scores 0.6, 0.2, 0.2. This would give cross entropy of 0.517209526. Once that is done for every item an average of these scores are calculated leading to an average cross entropy of 0.466682388 as provided in Table 34.

Table 34:

Step three in evaluating models using cross entropy.

Calculation of Cross Entropy for Each Name & Average Cross Entropy		
besme	0.517209526	
sena	0.359958891	
hadpl	0.522878745	
		Average: 0.466682388

The number does not have any meaning by itself. It only becomes meaningful when it is compared to a score of another model. For example, if the researcher decides to make changes to how the data is structured where instead of transcribing the data using sound segments, the addition of phonological or contextual features will be added. In other words, more information is given to the model as assistance which results in creating a new model. The researcher would want to know whether the addition of such attributes make the model better[47]. To evaluate the efficiency of the two models the average cross entropy is compared where (0 < better model < model) the superior model is the one with the score closest to zero.

100% success is not shown with entropy. Knowing whether the model got a perfect score cannot be shown as the score is not designed to show that. Instead KL is the metric which has a true zero. This is a manipulation of the score whereby true zero is designated as the actual data with 100%. Such a perfect score should never occur as it is a weakness of a model since performance errors are always a possibility.

4.5 Analogical Modelling AM

Skousen developed one of the two main computational analogy models that have been used for linguistic analogy formations. Since this thesis is a linguistic research, Skousen's model will be adopted because the other model, Tilburg Memory Based Learner, was designed from the start point as a computational algorithm from a computer science perspective

[47] Since in ay addition of attributes means more processing which is not something positive for a model.

(Daelemans & Bosch, 2005). Not only that but the two are very similar with almost equal success rates[48].

AM was written first in Pascal then Perl and recently the algorithm has been adopted in an interface-based software called WEKA[49]. The one used in this thesis uses an unsupported[50] Perl script which can be found in both GitHub and CPAN[51]. The differences between the one used here and from Skousen's program is in the way the data is transcribed and reported as will be explained. In addition, to such a change the guide files of the unsupported version are more in line with current machine learning programs than with Skousen's used terminology[52]. Other than that, they are all the same but there is a good chance that older versions won't work on new Macs.

AM Analysis Process

To use the model three files are required; a data file, a test file, and an AM script. The data file acts as the storage or an exemplar lexicon which stores the **analogues** and **base of analogues.** The test file is where the **base for new words** are given without the **new word**. The script is where the analogy algorithm takes the **bases for new words** in the test-file and searches for similar **base of analogues** and once they have been detected the **new word** will be given in the result file based on how the **base of analogues** and **analogues** behave.

Data File

Table 35 shows an example of a file modelling the CVCV hypocoristic pattern; a full example of a data file is given in the Appendix. Each line represents an exemplar, a name associated with the hypocoristic pattern. In the order from left to right, the first variable is the

[48] The third reason why AM was chosen is because it requires minimum background knowledge of coding.

[49] AM in the WEKA implementation is like a blackbox where one doesn't have full control of how it works. Not only that but after using the same data, in Perl and WEKA there was notable difference in the result leading to questioning whether AM works in WEKAin the same way it is with Perl.

[50] Stating that it is unsupported is because no literature references it and the webpage that has all of the earlier version of AM does not have it.

[51] https://github.com/garfieldnate/Algorithm-AM and http://search.cpan.org/~nglenn/Algorithm-AM-3.09/

[52] It is mentioned here because any discrepancy in the terms used might be due to having the literature of the program and Skousen's work be at hand in the explanation.

number of tokens. What 8 means is that there are 8 *bubu* in the data for the name *besme*. Following the number, a space is added and then the category of the hypocoristic pattern is given. It is the output of the name. This can also be rewritten using labels such as A and B[53]. Following the output, a space is inserted and the input or the name is given.

Table 35:

Example of CVCV data file.

```
8 C1uC1u  0CbVeCs==CmVe  besme
5 C2uC2u  0CsVeCnVe====   sene
4 C2uC2u  0ChVeCdVpCl==   hedpl
```

In a way, the name is structured like the weather data above in Table 28. Thus, in the first column a gender variable is designated. It can have three attributes 0 for female, 1 for male, and = for not available. After the first column, a consonant variable is given. It can be one of 29 attributes- 28 different consonants and a == which is null. After that is a vowel variable, then a consonant till the full name is written. At the end of the file is a comment. It does not have any effect on the algorithm. It is used to simplify the reading of the names in the data-file.

Table 36:

CVCV hypocoristic pattern data-file structure.

Token	Output	Input							comment
Frequency	**Pattern**	**Gender**	**C status**	**V Status**	**C status**	**V Status**	**C status**	**V Status**	
8	C1uC1u	0	Cb	Ve	Cs	==	Cm	Ve	besme
5	C2uC2u	0	Cs	Ve	Cn	Ve	==	==	sene
4	C2uC2u	0	Ch	Ve	Cd	Vp	Cl	==	hedpl

As an analogy to the weather data one can read line two of Table 36 as follows. If a name starts with a [0] for female as the gender status variable, followed by [s] as the

[53] Although this is not very practical since the researcher will have to go back and see what A or B means, it is the direction that was taken here because the text files which were used in the tutorials used this approach.

consonantal status variable, followed by [e] as the vocalic status variable, followed by a [n] as the consonantal status variable, followed by a [e] as the vocalic status variable, followed by a null as the consonantal status variable, followed by a null as the consonantal status variable, then it will have as an output pattern C^2uC^2u as its pattern status variable and it occurred 5 times. The weather table can be read similarly as, if a day has cloudy as its cloud status variable, not windy as its wind status variable, and October as its month variable then it will have as an output, raining, as its rain status variable which occurred once per year.

One other reason for using the AM in this study is because it requires minimum computational knowledge. However, there is a huge learning curve with structuring the data-file. The learning curve is more similar to a learning curve of learning to play the piano or paint a portrait. Royal Skousen, Deryle Lonsdale, and Dilworth B Parkinson (2002b) comments that "In summary, there are a few skills that contribute to the successful development of a dataset: the choice of the number of variables, identifying those features most relevant to the issues at hand, and being able to account for data instance differentiation. Being able to satisfy these desiderata is an art, and is best acquired through experience" (p.359).

Test File

The test file is almost the same. The difference is in not assigning a number and not having an output. Eventually the test file has the **base of new words**, and those items are what the model will be using to predict the output or **new words**. It is structured as shown in Table 37.

Table 37:

Example of test file with two items to test.

```
UNK  0CbVeCs==CmVe  besme
UNK  0CsVeCnVe====   sene
```

It starts with UNK which is stating that the output is unknown. Then it gives the name which should be in the same format as the data-file. In other words, it should include all features that is included in the data-file from gender to phonological features. Whatever form is chosen in the data-file, it should be mirrored in the test file. The final item written which is also optional is the name without the features. Here it is optional but not a comment. The program will use it as the designator for presenting the data. For that reason, it is better to keep it since it will help out in reading the result file.

AM Program

In short, the AM file is where the program is written. It takes the test file and data file and applies Skousen's algorithm to it and produces a result file where every item in the test file is given an output. One can treat the algorithm similarly to other data mining tools where the exact process is not important to know but the interpretation of the results should be learned[54]. However, a simple explanation as discussed in the literature will be provided.

In the case of the file used above, what the AM algorithm does is take the queried name, *sene*, and predicts the derivation behaviour of the name. For example, if the test file had only *sene* and the model's task is to predict which form will be used between C^1uC^1u or C^2uC^2u or C^1iC^1i, the algorithm will first search for all of the names that start with /s/ and group them together. Then will search for all names with /e/ and group them together. Then moves on to /n/ then /e/ then /se/ and then /sen/ and finally /sene/ until all possible groupings of all variables are considered. Each one of these subcontexts as Skousen terms will be taken as a group and analysed for disagreements.

Disagreement is like weight that is measured when it occurs in a subcontext. It occurs if not all of the names in the subcontexts has the same output. In other words, if the subcontext

[54] This is seen in many student projects that were done in the linguistic department at Brigham Young University. Here it is treated similarly where the interpretation of the model and its implementation is what is being concerned with. The exact way or statistical computational details that are used are not very important. This is also seen with how Skousen published AM where one textbook was designed for linguistic students and the other for computer scientists.

/se/ had in it the male name *saalim* and the female name *saalij* where the former has C^2uC^2u and the later had C^1uC^1u as the output then 1 disagreement would be tallied. Subcontexts with less disagreement will be left and used as the **base of the analogue**. In a way, it is similar to the weather model above where the similarities were tallied.

In the case of *sene* the group with less disagreement is chosen as the **base of the analogue** and in that group, are 3 other names which are interpreted as the ones closest to queried item. Thus, in that group, are *senaʔ*, *sula* and *sami*; whereby *senaʔ* and *sula* had the output C^1uC^1u while *sami* had C^2uC^2u. This can be summarised as in Table 38.

Table 38:

Calculations done in AM

analogue	base of the analogue	
C1uC1u (susu)	senaʔ	50%
C1uC1u (susu)	sula	40%
C2uC2u (mumu)	sami	10%

Now those are the details from which the result file is written. Interpreting them from here as to which is the winner or the pattern that *sene* will get can be done either by a selection of randomness. This is calculated as follows: since there are three possible outcomes for *sene* then each outcome has a 33% chance of being predicted. However, since out of those three outcomes two are similar, C1uC1u, then the probability of output C1uC1u is 66% (33+33) while C2uC2u is 33%. The second way that the results calculated is by calculating the most frequent out of the three and assign that as the winner. This is done by calculating the total of the probability of each analogue. In the example above it will be C1uC1u since 50+40 is 90 making it the winner. This latter interpretation is the one used with Models that predict winner as with Model A in Table 30. The former is with models that are like Model B in Table 30.

Result File

The result file that is produced will be found in the same directory where the three files exist. It includes details about which words in the database are similar to the test items which are the closest and which do the items form a link or a gang as Skousen terms it. For the objective of this study, the most important part of the result file are the results which are seen below in Figure 14. Skousen designed AM to give the two types of result mentioned above, a probability prediction and a winner[55]. Further, he designed it where 100% is not achieved because speakers do exhibit slips of tongue and other speech errors.

Figure 14:

Example of Result File.

```
Winners:
A
Scores normalized:
$VAR1 = {
          'C' => '0.0876216968011127',
          'G' => '0.19471488178025',
          'B' => '0.0987482614742698',
          'A' => '0.503940658321743',
          'H' => '0.00139082058414465',
          'K' => '0.113583681038479'
        };
Statistical Summary
+-------+---------+------------+
| Class | Score   | Percentage |
+-------+---------+------------+
| A     |  556544 | 50.394     |
| B     |  109056 |  9.875     |
| C     |   96768 |  8.762     |
| G     |  215040 | 19.471     |
| H     |    1536 |  0.139     |
| K     |  125440 | 11.358     |
+-------+---------+------------+
| Total | 1104384 |            |
+-------+---------+------------+
```

What the results show is that A is the winner. In other words, out of the different possible patterns that the word can have A is the right one. Below that is a statistical summary. In it are the probability scores discussed above. Hence there is a 50% possibility that it will be A, 9% that it will be B, 8% that it will be C, etc. These can be also interpreted as outputs that the model gave which is where linguistic variation can be modeled.

[55] In Skousen's tutorial on AM the choice of the result is made by changing the script of the program. However, the implementation used here doesn't give a choice and presents both types of result in every test.

4.6 AM used with Hypocoristics

In the rest of this study different models will be tested and evaluated. The –o affix will be used first to see how well the model can predict their three behaviours. Then, the CVCV pattern will be tested and four models will be used in the test and an evaluation of the three will be given. The simple metric will be used mostly in all data-bases as it is easier to interpret[56] than entropy which brings it closer to having a comparison with D&Z's approach. The CaCCuuC or default hypocoristics will be discussed in the next chapter.

-o Affix

To start with testing a model, a data-file is required. Since the –o affix is being modelled, a data file which is composed of all of the names that had the –o affix is created. As a starting point the name will be written without any specification as seen in model A in Table 39. It will just be the segments of the name and the output. In Model B, it will be just the consonants and vowels specified. In Table 39, Model C will have high vowels specified; this is highlighted with **H** next to the high vowel. This is done through the whole data.

Table 39:

Example of Three Data file structures for modelling -o affix.

Model A sounds only	Model B consonants and vowels only
5 C msa3ed msa3ed	5 C Cm==CsVaC3VeCd==== msa3ed
4 B fedwe= fedwe	4 B CfVeCd==CwVe====== fedwe
2 A hadi== hadi	2 A ChVaCdVi========== hadi
Model C specified H vowels	
5 C Cm===Cs=VaC3=VeCd=== msa3ed	
4 B Cf=VeCd===Cw=Ve===== fedwe	
2 A Ch=VaCd**H**Vi========== hadi	

[56] Due to the time limit of this study and the cumbersome effort that is taken in creating a model, testing it, and reporting it and in many instances fixing it, the simple metric will be used most as entropy is also more time consuming.

Since a token frequency is available and the size of the data is small, they will also be used as shown in the first number[57].The outputs will be given as A, B, C for [jo], [*Vo], and [#o], respectively as shown in rightmost column in Table 40. The second model will have consonant and vowel attributes added and a final model will have a [H] high front vowel that will be attributed since it is already mentioned that they are the factors for the different behaviour.

What should first be said here is that the analogical model in a usage-based theory contradicts itself if usage of such categories is used. In other words, a usage-based model should not require from formal linguistic details that describe detailed linguistic categories since categories in a usage-based model is defined statistically. In other words, a consonant is a segment because it appeared statistically a certain numebr of times before the other sound or after the other sound (Saffran, 2003; Saffran et al., 1996). Thus, having an analogy approach be successfully significant should only be taken when the first model which only has sound segments specified have a good success score. In this study, in all tests this has not been established. As a result, more tolerance for accommodating additional features will be taken in the approach.

Step two is having a test file. There are two ways to do that which are conceptually similar but different in the objective of the researcher. The first way which will not be done here is after a researcher has a good model. It will be used to predict unknown information. A great example is the weather model. Everyday millions check the weather forecast of the week. This is done based on an analogical algorithm similar to AM. In linguistics, one area this type

[57] . It is important to note that the tests done throughout were designed after many trials and changes. What will be reported will be the tests that showed the best results. The final choice of how to transcribe the data file as to which specifications to use and which names to leave out for the test file is what will be called manipulation of the data. It is manipulation in terms of the ability to play around the data but not in the sense of deception. The data is still as collected and has not been changed but way AM modeling works allows freedom for the user.

of modelling is used with, is language change and evolution where language is simulated (Eddington, 2004; Smith, 2014).

The other way is to split the data into a test file and a data file. In –o affix case, there are 128 out of 165 names that the –o suffix appeared. Thus, 128 will be split into a file for testing with 3 items and a file for data-file with 125. The reason for this split is to leave the data base as large as it can be since it is considered a small set. This is an example of what is meant my manipulating the data. However, the test will not be on just two names. There will be 2 tests and in each test 3 new names will be removed and the ones that were used will be returned.

The choice of the 6 items removed are based on the researcher's interest in investigating a certain area. For instance, will the model be able to find phonological patterns that were noticed even without ever specifying the data that a sound is a high vowel or not. The results of the testing are given in Table 40 and discussed below.

Table 40:

Testing three models with two tests using three names.

Names	model A	model B	model C	Actual choice	Actual Data
Group 1	**sound**	**Cons Vowel**	**High Vowel**		
ðˤaari	*C	A	A	A	ðˤaari-[jo]
rula	*C	*C	B	B	rul-[*Vo]
ʕabiir	C	C	C	C	ʕabiir-[#o]
Result	1/3=33%	2/3=66%	3/3=100%		
Group 2	**sound**	**Cons Vowel**	**High Vowel**		
miʃaari	*C	*C	A	A	miʃaari-[jo]
ðˤuħa	*C	*C	B	B	ðˤuħ-[*Vo]
riim	C	C	C	C	riim-[#o]
Result	1/3=33%	1/3=33%	3/3=100%		
Average score	33%	49%	100%		

In model A where names were written without ever specifying the segments, the model did poorly. Only 2 out of the six were predicted correctly. As a result, another model is created where consonants and vowels are designated. The model did better, as shown in Table 40, with only 3 correct items.

Finally, the data included specification of vowels. Once they were included the model performed perfectly with all of the data being predicted correctly. A final test was done with the same high vowel specified but this time with all 6 names at once meaning that it had less data to search through. The model still worked perfectly predicting the correct pattern for every name.

The scores above are then used to interpret them as the researcher wants to. For example, one interpretation is that the high vowels play an effect making them a more plausible psychologically represented category. Another finding can be seen with the failure of a test. Model A can be seen as a child lexicon where few items exist. The results taken here will be compared to child data and if correlations are found then it shows that failure of the model and wrong usage of the –o affix with child are the result of having a small lexicon.

For example, after looking at one specific child which contributed to the collected child data; out of 7 names that the child was given, zero were wrong. Now it is expected that the child will not know 125 names. Hence the child outperformed the model which used a database with 125 names. Such finding is how various computational models are used. Overall, when compared to three phonological rules that apply after adding the suffix as with generative theories, analogy show that it is incompatible with the data and reasons won't be given and known due to lack of a having a larger data base or corpus where further tests can be made.

CVCV Pattern

After the –o suffix test was completed, five different models were created to test the CVCV data. Again, in Model 1 the sounds without any attributes were used. Model 2 had consonantal and vocalic features used. In Model 3, only the gutturals were specified. In Model 4 both gutturals and glides were specified. Finally, the model with the best score was taken and the gender attribute was removed from the data resulting in model 5. This is done to see whether such a category has an effect especially since there is an observation that gender of the name plays a role with *susu* where it showed mostly wit female names and not male names despite having the male name start with an /s/.

Again 6 names are chosen in the test file and two tests were done with 3 names each. The table will have the full hypocoristic specified because unlike previous data there are 5 patterns. The data with the CVCV hypocoristic is different than the –o suffix. It is usually with similar type of variation that the scores are given using entropy. Here, it will only be used with the final data and instead the same approach test that was used in the previous chapter to evaluate D&Z's rule will be used. In other words, if the model predicts a hypocoristic that can be found in the data then it will be taken as a correct prediction. If the prediction is not chosen by subjects, then an asterisk will be added, and it will lower the score of success. Furthermore, the names that will be chosen are those that have avoided the most productive CVCV pattern which used the initial consonant. The purpose of the test is to check whether gutturals are avoided without specifying them and whether back glides are avoided with specification as opposed to front glide. Since front glides are accepted will a specification of just glides to both front and back glides be useful.

	model 1	model 2	model 3	model 4	model 5
Group 1	sounds	+CV	+CV+G	+CV+G+Y	Best -gender
ʔasˤaajil	*ʔuʔu	*sˤusˤu	*sˤusˤu	*ʔuʔu	*sˤusˤu
saami	*susu	*susu	*susu	*susu	mumu
leen	lulu	lulu	lulu	lulu	lulu
Result	1/3=33%	1/3=33%	1/3=33%	1/3=33%	1/3=33%
Group 2	sounds	+CV	+CV+G	+CV+G+Y	Best -gender
ʕafaaf	*ʕuʕu	*ʕuʕu	fufu	*ʕuʕu	fufu
suʕaad	susu	susu	susu	susu	susu
wafaaʔ	*wuwu	*wuwu	*wuwu	*wuwu	*wuwu
Result	1/3=33%	1/3=33%	2/3=66%	1/3=33%	2/3=66%
average score	**33%**	**33%**	**49%**	**33%**	**49%**

The table can be summarised as follows: Model 1, Model 3 and Model 4 had the lowest scores even though Model 4 had gutturals, consonants, vowels, and glides specified. Model 2 with only the consonants specified did better than model 4. Model 3 and 5 predicted the most productive patterns. These are the models where gutturals were specified in addition to consonants and gender for only Model 5.

The three most noticeable behaviour is seen in the table can be summarised as follows. Specifying gutturals in models 3 and 5 led to having the model not violate the constraint. Despite avoiding gutturals, model 3 and 5 did not avoid *sˁusˁu*. Finally, adding gender feature showed that the model can predict lack of *susu* with male names whereby *mumu* was chosen instead for *susu* for *saami*.

There were three complications that cannot be explained. First, the violation of back glide with homorganic vowel is consistent as seen in *wuwuw*, even after the glide was specified. Second, after specifying consonant and vowels, the model avoided a constraint against the glottal stop and chose *sˁusˁu*. Third what is still not understood is why adding more specification to the model impairs it as seen when glide was added.

The result can be used to show that having a constraint in initial gutturals is clearly valid. What is also interesting is that despite having specifying gutturals *sˁusˁu* appeared in the data showing a similar the tug of war which Prunet described between functional and formal analysis. *sˁusˁu* has a high token frequency leading it to violate any structural constraint against it. In the gender case this frequency also played a role by having *mumu* instead of *susu* where the data had a low frequency of *susu* when used with male names.

4.7 Short Discussion and General Findings

In general, what can be noticed is that adding more specifications can aid in the prediction however this doesn’t always occur. Another thing noticed is that the model is sensitive to minor changes which is a feature that token based models and exemplars have. In the first chapter, it was mentioned that having Generative grammar be rule-based aids in describing a language. This is seen here where if asked to give a person the grammar of the usage of the –o suffix it will only be achieved by generative formations. A table with probability of analogue occurrences of similar forms is not a grammar in the traditional prescriptive sense.

It is important to stress here that one can interpret the addition of guttural as support for an argument for the psychological reality of the category, including gender of the name where *susu* did not appear with male name despite it being phonologically valid. At the same time this could simply be that AM simply performs better with more details. Thus, more work is required as the results aren’t conclusive

The chapter started with OED’s definition of analogy which was argued that it was vague for not having specification on what is used in an analogy process. The chapter attempted to remove the vagueness by being specific in how an analogy can be applied. Specifically, it used a computational analogy model which is more rigid and less vague since precision of the calculations are supposedly consistent. This would still be the case with any computational model as they are specific since they are based on constant calculations. If the analogy model chosen was one that is not computational, the various questions of determining the **analogue** will be left unanswered.

5. Discussion

Skousen (1989) states that in many cases the predicted behaviour is nearly the same no matter whether a rule approach or an analogical one is used, but conceptually the two approaches are vastly different. The conceptual differences between the two have put them in competing positions. In this chapter, an evaluation will be provided where the weaknesses of both, AM and D&Z's analysis will be discussed. This will lead to presenting a third approach showing both a rule-based and analogy-based hybrid. The hybrid approach will be presented after showing other hybrids that have been proposed in literature.

5.1 Weakness of AM

There are two noticeable weaknesses of AM. First, is the human interaction aspect that leaves openness in the way a model can be tested. This lack of restriction contradicts the general purpose of a computational analogy where formations are result of a specific calculation. It also returns the arguments put forth against analogy for not being able to be specific in the choice of which items are used for the analogy process (Blevins & Blevins, 2009; Itkonen, 2005). The second noticeable weakness appears with how the model behaves with small data-sets; which can also be equated to a child lexicon.

Human Assistance and Result Manipulation

One area that analogy has been criticised is with how an analogy relation is not restricted (Blevins & Blevins, 2009; Itkonen, 2005). In the analogy, 2 is to 4 as 3 is to 6, there is a relation that exists in 2:4 leading it to have a correspondence analogy relation with 3:6. The relation is the multiplication of 2. However, there can also be another relation between 2:4 which is the addition of 2. In this case, the corresponding pair will be 3:5 leading the **new item** to be in another form. This wide scope of analogy where multiple relations could exist is claimed to not be available in computational analogy which is an advantage of computational analogies (Eddington, 2004). The algorithm of AM computes the relation in a single consistent way with every analogy process resulting in specificity of the process. In other words, the search for the **base of analogue** and **analogue** for which the **new item** is chosen will always be the same.

From a practical position this is not the case. There are noticeable aspects of human interference where the precision of a computational analogy can be manipulated. One example comes from the way certain non-contrastive sounds are transcribed[58]. Transcribing the name *ʕalij* as *3alp* where /p/ is one segment used for a long vowel or *3aliy* where /iy/ is two segments is one that was repeatedly causing issues in the prediction results. By treating it as a short vowel with a glide /iy/ it was influenced by names with glides; whereby glides are found in many names, and since the algorithm will take every segment and use it in the search of differences, *fe**j***, *ʃa**j**maʔ*, and *miʃar**ij*** would all be considered to play a role in probability of their usage. In earlier tests, this issue lead to having [–o] pattern suffix appearing after *3alp* but not *3aliy*. For that reason, the data transcription was manipulated constantly by having long vowels be either one segment or two segments till a high prediction score was reached which was reported in the previous chapter. Again, the manipulation should not be taken with a sense of being deceptive approach by the researcher but instead it can be seen either as a weakness of the method or as Skousen and others called a form of art that is part of how one uses the algorithm.

The human aspect of creating a model also appeared on how the test was structured and what names are used in a test. If a test had more than one name with similar structure, a change in the results was noticed. In addition to the choice of the names, the number of names used in a test had an effect on the outcome. In the reported data, these two were factors that were purposely manipulated to get the best score especially since the data-base is not large[59].

[58] Although long and short vowels are in contrastive distribution in Arabic, with names a contrast is not noticed. More work is needed to show how this behavior holds up in Arabic. However, it has been shown that the phonology of proper names is different from that of nouns (Brennen, 1993).

[59] Postscript: it has been brought to my attention after the tasks were done, that the WEKA test-task is designed without access to the control of the split of the data. AM in WEKA takes one item and tests it, then another, then another till all of the items are contrasted against the actual data-set and an average entropy or another evaluation metric is used to evaluate the model.

In earlier trials not mentioned here, the results were very weak. For that reason, a trial and error process had to take place to achieve the high result reported. Whether it is AM or another computational model, the need for human assistance is required in an analysis. The assistance of a model in compiling and operating it is not a weakness of computational approaches in general. However, the assistance of setting what categories to look for and how to set them or what to include in AM, opens area for flaws and criticism from a computational, linguistic, and a scientific view. It is important to note that after a while of using the AM program, a researcher would have the ability to direct the model into whatever result is required. A great quote from one of the contributors of AM is given below, which captures the criticism that is to be conveyed on the manipulation aspect of AM as a computational model:

> A number of questions arise regarding the selection of database items and variable selection. Is a database of 939 items too many or too few? Is it a fair approximation of what Spanish speakers know? This database contains types and not tokens. Would a database generated on the basis of token frequency be more representative? The variables in the database represent phonemes and are organized according to syllables. Would it be better to consider phonetic features or acoustic qualities rather than phonemes? Perhaps some alignment of the variables other than according to syllables would be more psychologically plausible. All of these are valid questions that have yet to be answered. (Eddington, 2004, p. 84)

Lonsdale’s description, quoted earlier, of the process in his AM tutorial on how to set the data, describes the process as an art; it is a great analogy to point. This would conclude in two things, first questioning the empirical validity of computational models that have an artistic element to it. Second, questioning the specificity of any computational analogy algorithms which would entail a weakness in the directions that are taken today with analogy as used in computational models (Blevins & Blevins, 2009; Itkonen, 2005).

Child language

Considering how the number of items in a database makes a huge impact on the result is a weakness of AM as mentioned above. In the children’s hypocoristic collection, what was noticed is that a child knows how to derive hypocoristics like the adults at the age of 6[60]. Hypothetically[61], how many names has that child come across? The answer will be the same number of people the child encountered or have seen on TV. In addition to that, names come in a trend (Pinker, 2008). So, at a certain period of time there will not be many varieties of names especially in a small culture as Kuwait.

The hypothetical argument above is written to simply state that children store few names. In relation to an exemplar lexicon this would mean that the data-base of a computational model would contain few names. As seen, with a small data-base, low success rate is achieved as with the –o suffix model. On the other hand, at age 6 and 7 children show fluency with hypocoristics with outputs that are the same as the adults with no mistakes. This results in not being able to claim full success of AM as an analysis approach for hypocoristics.

5.2 Weakness in D&Z’s Approach

Hypocoristics are similar cross linguistically (Brylla, 2016). They are formed in every language referenced here, either with phonological rules or the concatenation of affixes. D&Z’s analysis resulted in having Arabic hypocoristic formation be an exception. Their claim that hypocoristics are formed morphologically entails having names be semantically decomposed; even though names cannot have a decompositional semantic meaning.

Proper names are semantically opaque acting like blocks which cannot be broken up. The surnames *Baker* and *Letterman* are not *bake* + *er* or *letter* + *man*. The only way the meaning is known is by looking at a name dictionary or an encyclopaedia. This occurs in every

[60] The children data was not included or discussed throughout the study because it is similar to adults.

[61] The argument is hypothetical and cannot be supported since research on vocabulary counts exclude proper names (Nation & Waring, 1997); and no work has been come across from onomastics or language acquisition that shows that.

language that is referenced here, and it is a topic that onomasticians have dealt with (Brylla, 2016; Leibring, 2016; Van Langendonck, 2007; Vom Bruck & Bodenhorn, 2006). For that reason, D&Z’s analysis is rejected[62].

Meaningless Hypothesis

Throughout chapter three the phrase **smallest reoccurring formative that carries a distinct form and meaning** was repeated and highlighted various times. The claim was that any formative that is given a place in the lexicon should have a reoccurring meaning with a single form. With regards to D&Z’s approach, this would imply that speakers are decomposing the name and extracting a formative, C-root, and mapping it on a template[63]. In addition to that, the meaning of the C-root is known to the speaker. This is why their approach is not only unique but unattainable.

Proper names are formed or given at first with a semantic or any associative meaning. Later the meaning gets lost, either due to language change, hypocoristic formation, or the change of status of a name where the original meaning associated with the name is detached (Leibring, 2016; Vom Bruck & Bodenhorn, 2006). In addition to the many onomastic work that show impossibility of this occurring, various evidence from different fields point to names being meaningless or what has been known as the meaningless hypothesis, Millan Approach to names, or Baker/baker paradox.

This complication was first noted in (Idrissi et al., 2008) (henceforth IPR). IPR did not argue against the existence of a C-root nor that the proper name is the input from which the hypocoristics are formed. Their argument was against having proper names being subjected to morphological decomposition. Again, following Structuralists as well as Generativists that hold the Lexicalism hypothesis, a morpheme is a sign that has a reoccurring form and

[62] Other reasons include having hypocoristics for names with no roots, and having multiple variants for names that do not follow a C-root.

[63] In their first papers they clearly show this but later published that the c-root is referenced.

reoccurring lexical meaning. The existence of both is required for any morphological decomposition as Aronoff pointed-out with the problematic cranmorphs. In other words, for one to decompose the word *reuse* to *re* and *use*, both *re* and *use* are required to occur in other environments with the same meaning. This is what is meant by a reoccurring meaning.

Thus, the *re* in *replay* and *review* can be decomposed into a separate morpheme because it shares the same form and meaning of ‘to do again’. However, in the word *recent* it is just a syllable that happens to have a similar form but not the same meaning. By not sharing the same meaning it is not a morpheme even if the other part of the word *cent* does share the same form as the word for ‘sum of a money’.

In relation to proper names, the search and decomposition of names into further morphemes using the reoccurrence of meaning is not available; due to having names be meaningless. There is a general consensus going back to Ancient Greek philosophers that the semantics of proper names is different from that of common nouns (Summerell, 1995). This continued even amongst modern day semanticists and philosophers with the view that solving the question on the difference between the two, would solve a “problem of meaning and reference” (Van Langendonck, 2007, p.22). For that reason, there are more than five theories concerned with the semantics of proper names (Vom Bruck & Bodenhorn, 2006).

From a Generative perspective, the problem lies in the representation of the units stored in the Lexicon. In addition to phonological information, semantic information is also stored[64]. Whatever form it is stored in, it can be described by its sense. The sense of a morpheme is the concept or underlying meaning that is denoted. For example, the sense of *un*, when used with adjectives is the ‘absence of the adjective’. In a way, it is similar to general descriptions given as dictionary definitions.

[64] The question of how it is stored is not a topic that will be addressed since it is a semantic issue that will require another thesis to investigate it.

Contrasted with the sense of *un* or *play*, the sense of proper names is referential[65]. The name *John* has as its sense someone to refer to. One might reject this by saying that every common noun is referential. For example, the sense of *dog* is a referent of the animal also known as canine. This argument is falsified by contrasting the types of referents.

First, the referent of proper names denotes a constant referent that varies amongst speakers and contexts. They are similar to indexicals. In other words, the meaning of *John* for speaker 1 can be ‘the neighbour’ and for speaker 2 it is ‘one of the teachers’. On the other hand, the referent of dog might vary within a few speakers but in the general sense it has underlyingly a single conceptual referent that would act as a categorical hypernym. This leads to the second difference; the referent of John cannot be ‘all of the Johns that exist, existed, and will exist’. However, the referent of dog can.

Having proper names carry a referential sense that is never consistent amongst speakers entails that it cannot have a reoccurring meaning listed in the lexicon. As a result, any attempt of morphological decomposition like the one D&Z presented are invalid. The basic principle of decomposition, which searches for a reoccurring meaning is not available.

This concept is also behind the solution for the Baker/baker paradox in psycholinguistics (Bonin, 2003; Brédart, 1993). Behaviour differences between the processing of the name *Baker* and the noun *baker* occurs due to having the latter be a complex word with a lexical meaning as opposed to its homonymous name. In online tasks, subjects show a preference for memorising a person who is a baker but not a person whose name is Baker. The reason is due to having the aid of a semantic memory which proper names lack; leading to the meaningless hypothesis. This is the main reason for IPB’s rejection of D&Z’s analysis and why it is also seen here as the weakest of the two approaches; other being AM.

[65]The approach taken to show the difference is one of many methods used in Semantic Theory to distinguish between the two.

5.3 IPB

If semantics, which is a crucial element of morphemes in Generative Theory, cannot be used, one would assume that the result is a derivational process almost similar to that of Sonority Stripping or Prosodic Morphology. In other words, without semantics the only possible way to derive a hypocoristic from a given name is by specifying the segmental material in the proper name phonologically. Yet this is not what IPB proposed (Idrissi et al., 2008; Prunet & Idrissi, 2014). IPB proposed a hybrid model.

Hybrid Model

One topic that has been repeated in this thesis is the existence of theoretical positions that are on opposite ends. The rule-based versus analogy-based, functional versus formal, word-based versus morpheme-based, and performance versus competence positions are drawn here as a rigid dichotomy. However, a closer investigation reveals that a continuum exists, and the positions presented thus far in this thesis took the opposite ends of the dichotomy. The best way to show these dichotomies is with two hybrid models which fall in the middle.

In many of his published work pinker pointed out to the deficiency of both generative and analogy approaches (Pinker, 1999). While both do have areas that are in line with linguistic evidence, the place where they fall short is with what each treat as irregular. As such he devised a theory which uses rules with the regular formation and expands an analogy formation with irregulars.

What Pinker fails to show is a precise way of regulating the two systems. He doesn’t provide with the answer to when are lexical items formed via analogy and when are formed by rules. Some successful hybrids resort to such precise restriction. J. C. Watson (2006) is one approach that does that in the dichotomy of a WB and MB lexicon. Using San’ani Arabic diminutives, she introduced a framework that is “neither an entirely root-based nor an entirely

[...][word]-based approach, rather claiming [...] that both types of word formation occur in Arabic" (J. C. Watson, 2006, p. 190).

Having the two approaches creates competing derivational systems where it is usually the case that a delimiting tool is posed to determine how one type of analysis works while the other doesn't (Aronoff & Lindsay, 2016). Watson's approach for the competition was done by restricting MB derivations to diminutive verbs and MB derivations to the nominals. For example, after she elicited speaker's meaning for a set of verbal diminutives she saw that a form and function relation can be recognised (as seen in Table 41) based on a shared C-root which "suggests both that the basic consonants are extractable from the tCayCaC form and that the triliteral consonantal root is recognised by speakers as an independent morphological unit" (J. C. Watson, 2006, p. 193). On the other hand, she noticed a few words that do not comply with a C-root derivation and instead suggest that they are derived from words. For example, due to the initial bilabial the diminutive *tmaydar*, is assumed to be derived from *mudiir* and not the C-root *d-w-r* which lack the bilabial segment.

Table 41:

Verbal Dimnunitives

	Verbal Diminutives	Elicited Words	C-root	C-root Meaning
1.	txaybal	xabal, mixgaalih	x-b-l	relating to stupidity
2.	tlaygen	lagaanih, layganih	l-d-n	relating to bicker
3.	txaydaʕ	xadaaʕah, xadaaʕ	x-d-ʕ	relating to deception

IPB's Hybrid Model

IPB proposed that hypocoristics are derived from "names solely on the basis of their surface form" (Idrissi et al., 2008, p.246). Thus, it is similar to Ratcliffe's approach in which it is word-based. However, the derivational process is guided by the saliency of the root-and-template morphology. To them, the evidence from psycholinguistic studies on Semitic languages show that the root-and-template morphology is "rich enough to prompt speakers to extract roots in the absence of semantic cues" (Idrissi et al., 2008, p.246). Thus, it also incorporates elements from a MB lexicon making it a hybrid between a WB an MB approach.

IPB are proposing the following. A speaker would learn the root-and-template morphological derivation that richly exists in other parts of the language such as plural and dual formation, or nominalisation. Then, when faced with a derivational domain that is free from lexical C-roots such as loan word adaptations and hypocoristics, the speaker would still use the same method but with form alone. For example, because in a XVCVCVC derivation the first three Cs following an X are usually the C-roots in a language, then in a loan word with the same structure the first three Cs would be adapted into the language and used for the derivation. Note, it is important that the process is linked to having a MB root-and-template derivational system be psychologically represented in order to guide through the derivation. It does not side with Ratcliffe's proposal against a MB lexicon nor does it side with D&Z to have the lexical C-root be the unit used in the derivation. It is like Pinker and Watson's hybrid approaches in terms of where it falls theoretically between WB and MB.

In addition to that it falls between functional and formal theories. In their work[66] they clearly state that functional investigation is required due to stochastic observations in hypocoristic formations but needs further investigation to prove empirically (Prunet & Idrissi, 2014). This is where their claim can be understood when they stated that "no linguist can speak for the community when it comes to the analysis of names, since we are dealing with extralinguistic competence" (Idrissi et al., 2008, p. 247). In a way, their approach is similar to Chomsky's call for a generative Morphology. Thus, the call for IPB's (Henceforth Prunet[67])

[66] IPB have given various papers and presentations on the topic. In addition to what has been published many correspondences have occurred while writing this study where their work was questioned for a better understanding.

[67] Most of the correspondences and conferences that their work was presented in was done by Prunet, and as such what is said references Prunet as an accumulative of all the material presented here.

approach with hypocoristics will be given below and it will be shown how what they were claiming is better seen as an analogy approach.

Prunet's Cookie Cutter Theory

What Prunet is claiming is that a speaker would have a MB lexicon where C-root and template formation are active in domains such as plural formation, passive formation, causative formation, etc. (Table 42). Now in these actively C-root formations, the C-roots have designated position that alter their position in the derivation process. For example, the plural noun *ʔaʃʤaar* 'trees' falls in the following paradigmatic forms *ʃaʤara*, *taʃʤiir*, *juʃaʤir,* and *ʃaʤarataan*. Thus, the constant root which is used in the paradigm falls in second, third and fourth consonantal position of *ʔaʃʤar* and is therefore the C-root. This appears with other words that have ʔaCCaaC structure. Thus, in names that have a similar form as *ʔaʃʤaar* like *ʔaʃraaħ* the speaker would treat the consonants positioned in these places as the C-roots and map them on to the hypocoristic template CaCCuuC.

Table 42:

Arabic Acitve paradigm relation

	plural	singular	passive	present	dual	C-root
1	ʔaʃʤaar	ʃaʤara	taʃʤiir	juʃaʤir	ʃaʤarataan	ʃ-ʤ-r
2	ʔafkaar	fikra	tafkiir	jufakir	fikrataan	f-k-r

Furthermore, since names have various forms, a speaker would form various rules that are dependent on the formations of the name (Table 43). For instance, for the production of the following names, *baasil*, *ħamdaan*, *ʔaħmad*, *marjam*, and *ʔibrahim* the speaker would require some of the rules below, based on having these structures existing in the language as C-root rules (see Appendix for full list of rules and names).

Table 43:

Rules used for default pattern

#	Category	Pattern			name	hypocoristic
1	L	C1V0C2V0C3V	→	C1aC2C2uuC3	baasil	bassuul
2	G	C1V0C2V0C3V0aan	→	C1aC2C2uuC3	ħamdaan	ħammuud
3	H	ʔaC1C2V0C3	→	C1aC2C2uuC3	ʔaħmad	ħammuud
4	I	C1V0C2C3V0C4	→	C1aC2C2uuC3	marjam	marjuum
5	J	ʔiC1tV0C2V0C3V	→	C1aC2C2uuC3	ʔibrahim	barhuum
6	K	ʔiC1tV0C2V0C3V	→	C1aC2C2uuC3	ʔbtisaam	bassuum
7	M	C1V0C2V0C3aʔ	→	C1aC2C2uuC3	ʔasmaaʔ	ʔassuum
8	N	C1V0C2V0jV0C3	→	C1aC2C2uuC3	ʔasajil	ʔassuul
9	O	C1V0C2V0wV0	→	C1aC2C2uuj	salwa	salluuj
10	P	mV0C1V0C2V0C3	→	C1aC2C2uuC3	muhanad	hannuud
11	Q	C1V0C2V0	→	C1aC2C2uuj	muna	mannuuj
12	R	C1V0C2V	→	C1ajjuuC2	liina	lajjuun
13	S	C1V0C2	→	C1ajjuun	faj	fajjuun
14	T	C1aC2aaC3iC4	→	C1aC2uuC4	ʤawaahir	ʤahhuur

The problem with such an approach in a Generative Grammar is with the formation of these rules. First, in a generative framework, rules are context free. Specific phonological or lexical material do not belong in grammar rules. A rule such as D&Z’s, X → Y where X is the C-root of the name and Y is the hypocoristic template CaCCuuC, neither carries lexical detail nor phonological material. It can be a rule that is part of the initial state/Universal Grammar while the specific C-root and template are stored in the lexicon. On the other hand, Prunet’s approach is dependent on the shape of the name and thus leads to various rules targeting specific items in the lexicon.

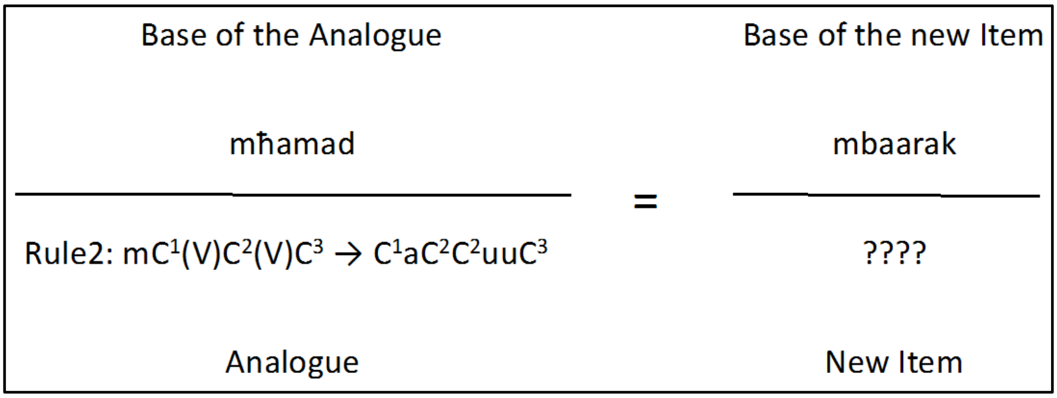
Second, in a rule such as rule 11, where it will be used with the names *mbarak* and *mħamad*. The rule would also be used with *marjam* which will produce the wrong hypocoristic. How will it not be used with *marjam*? This issue will only be solved by adding more rules where *marjam* will have a rule (rule 4), and *mħamad* and *mbarak* will have a different rule (rule 11). By adding more rules like these, they become non-productive since they apply to one item in certain instances; and as a result, is not applicable in a generative paradigm.

Hybrid Model Testing

Nonetheless, if Prunet’s approach was drawn similarly as an exemplar theory that uses analogy then what they have presented could be modelled using Skousen's AM. If every name has a pattern defined over its structure and an associated output, this will be exactly like an analogy approach. In various presentations that Prunet presented he stated that hypocoristic formation is like a derivation that used a “cookie cutter”. A person comes across a new name, and then searches for a name with similar structure from a bag and then uses the name as a cookie cutter. This is illustrated below Figure 15.

Figure 15:

Rule-based-analogy theory.



A speaker comes across a new name, *mbarak*. The speaker will want to form a hypocoristic for it. A searching process looks at similar names. Once found, the hypocoristic rule that is associated with the name would be applied to form the new hypocoristic. What has been described has been modelled in AM.

After, creating three models, one with only phonemes, one with consonants and vowels specified, and one with glides specified, an average entropy comparison was taken and shown below in Table 44. The comparison shows that no significant difference is seen between the CV Specified model and when it was specified with glides. However, the phonemic model performs the worst significantly[68]. Therefore, the CV-model with specified glides will be taken to show how the outputs compare to actual data.

Table 44:

Comparison between cross entropy of three models

	Phonemic Model	CV Specified Model	CV Specified + glides Model
cross entropy	5.154	0.963	0.756

A note needs to be said here about the structure of the data-file. First the file can be structured differently even with rules. Second is that 17 names were problematic. They either did not have a hypocoristic provided or they will have a rule just for them. For that reason, they will not be included and are listed in Table 45.

Table 45:

Problematic Names for Data-File of AM.

#	Name	Hypocoristic	Reason for not being modeled
1.	ʔanwar		No Hypocoristics
2.	ʕiisa		
3.	dawuud		
4.	foz		
5.	musˤtˤafa		
6.	nawaal		
7.	nuuħ		
8.	nuura		
9.	nuurija		
10.	nɔf		
11	ʔaħlaam		
12.	jaħja		
13.	badrija	baduur	requires a single rule
14.	muntaha	mantuuj	
15.	nawaaf	najjuuf	
16.	suhajla	sahhuul	
17.	ʔesmaaʔ	sammuuj	

[68] This number might be due to an error to the way the nulls were added in the data.

As explained in the previous chapter the model results are given as winner and as probabilities of each output. What will be reported are the winner and the second rule with highest probability. Thus, Table 46 will show which rule category will be predicted. If the correct rule is predicted the full hypocoristic will be written. If a wrong rule is predicted the category of the rule will be listed. The test will be a 1-fold test where 15 names will be tested one by one. The names include ones that D&Z predicted correctly, did not predict, and predicted wrongly. This can be compared in a similar fashion to D&Z's evaluation.

Table 46:

Comparison between Actual Data and New model output

#	Name	Actual hypocoristics rule	Winner	2nd rule
1.	daana	dajjuun & dannuuj	dannuuj	dajjuun
2.	leen	lajjuun & lannuuj	lannuuj	lajjuun
3.	raaʔid	raʔʔuud	raʔʔuud	H
4.	msaaʕad	masʕuud & saʕʕuud	masʕuud	P
5.	ʔasmaaʔ	ʔassuum & sammuuj	L	ʔassuum
6.	sanaʔ	sannuuʔ	sannuuʔ	H
7.	ðˤuħa	ðˤaħuuj	L	ðˤaħuuj
8.	miiʕaad	maʕʕuud	maʕʕuud	P
9.	wafaʔ	waffuuʔ	waffuuʔ	H
10.	mufiida	fadduuj	L	fadduuj
11.	yosif	yassuuf	yassuuf	-
12.	marjam	marjjuum	marjjuum	P
13.	saara	sarruun	Q	S
14.	zein	zannuj & zajjun	L	zajjun
15.	ħsein	ħassuun	ħassuun	-

The result of the model can be summarised as follows. Out of the 15 names the model was able to predict 16 correct hypocoristics. However, out of those 16, only 10 were the ones that were winners. For example, *marjam* had 2 rules apply to it, the correct rule deriving *marjjuum* which was the winner and rule P which was the second predicted rule and was incorrect. There are small dashes to show that the model did not predict another rule.

Overall the model predicted 12 wrong rules and 18 correct rules. There were some interesting predictions that were shown in the model that can be compared to D&Z's theory.

First, the names 1 and 2, had two hypocoristics in the actual data. D&Z’s rule could predict only one of them while the model predicted the two. Second, names that are Hebrew were not predicted in D&Z’s model. Here they were perfectly predicted as in 11. The most chosen rule was L, which is the most productive. This is probably due to its frequency.

Difference between Prunet’s Hybrid Model and Skousen’s AM

A question arises with what the difference would be if the traditional way of organizing a data-file was used instead of the approach taken in the hybrid model. In AM, data-files are organized by having an input associated with a single output. For example, if English past tense is being modelled, the data file would have present verbs as the input and the three ways of forming past tense as an output listed as three categories. A for /d/ after voiced, B for /t/ after voiceless, and C for /id/ after dentals. This is similar to the –o suffix and CVCV hypocoristic.

If this is followed with the default hypocoristic, then there will be one category since all of the hypocoristics have the CaCCuuC structure except for the names categorised in 12-15 of Table 43; resulting in 5 categories. Thus, there will be various details of how the hypocoristics are formed which are captured in the hybrid model. This leads to concluding that it is better to model the default hypocoristic similar to how it is given in Table 43 making it equivalent of any Skousen Model.

5.4 Summary

This chapter was the discussion chapter of this thesis where an evaluation of the approaches was provided. The points that were discussed are weaknesses of both Skousen’s Analogical approach and D&Z’s rule-approach. Skousen’s approach has weakness in the computational side of the program. The way it behaves with small databases and the freedom one has in manipulating the data all leave AM with a wide area of criticism that analogy has always faced, which is being not very specific and detailed.

The discussion continued with the main and only weakness of D&Z’s approach which is semantics. A root-and-template approach is not valid with names because names are meaningless. The criticism to D&Z’s analysis which is argued in various fields and areas comes from the meaningless hypothesis.

In return Prunet presented another view as a response to D&Z’s complication with their approach. The approach doesn’t belong in a rule-based grammar. It does use rules but the way they are chosen can be seen as an analogy. For that reason, an attempt followed which placed the proposed analysis within Skousen’s analogy approach. The approach was modelled and tested with good results. Further, the new hybrid model differs from Skousen’s AM in how the data is structured. However, it seems to be a better way due to complications with hypocoristics. In the following chapter, some aspects of hypocoristics will be shown which argues for having hypocoristics be analysed under a usage-based theory where frequency and contextual usage play a role.

5.5 Thesis Question

In the first chapter the main question and the answer of the thesis was presented regarding the status of hypocoristics and whether they can be analysed. The given answer was that **hypocoristics do show unpredictable patterns but depending on the theoretical approach taken these patterns can be reduced leading to a conclusion that hypocoristics are not extra-grammatical and an analysis is possible.** D&Z, Ratcliffe, and Prunet provided an analysis approach that used rules. Each one of these approaches showed potential of having a rule-system that can account for the derivations. However, the first two fall short in providing with details on semantics of names and detailed approach, respectively. It is only Prunet’s approach that is complete. However, the way it is given is confined to the research paradigm of formal linguistics with argument for the requirement of further functional research.

AM is a functional paradigm, as a computational model it can show success, but this success can easily be interpreted as failure. The easiness for manipulation of the data leaves a researcher with the ability to show success or failure with experience. This should then be taken as a failure of the approach from a computational side. One of those tweaks that were made was having the output or analogue representing rules instead of items. This hasn't been done before computationally. However, the success from such an implementation directs; first towards the search for a computational model that underlies a rule-based derivational system. Second, it points to having such a system be part of a usage-based theory instead of a generative paradigm. This is not only supported by the model that was presented but by some general findings from onomastics that can be shown with frequency effects. This will also benefit language description for an analogy approach.

5.6 Conclusion

The objective of having an analysis for hypocoristics has concluded. A rule-based approach has its descriptive merits. An analogy-based approach also has its merits. A question is then left with which can be used for further research on hypocoristics. What has been shown is an incline for future work for both approaches, mainly because both there are areas that require more explicitness.

Appendices

A.1. Symbols Used for AM Perl Data File

consonants			vowels			gender		sounds category		Affix class		Natural Class	
ʔ	أ	2	i	◌ِ	i	male	0	consonant	C	root	R	Gutturals	G
b	ب	b	ii	يِي	p	female	1	vowel	V			Glides	Y
t	ت	t	u	◌ُ	u								
θ	ث	8	uu	وو	o								
ʤ	ج	j	a	اا	a								
ħ	ح	7	e	◌َ	e								
x	خ	5											
d	د	d											
ð	ذ	v											
r	ر	r											
z	ز	z											
s	س	s											
ʃ	ش	c											
sˤ	ص	9											
ðˤ	ض	x											
tˤ	ط	6											
ðˤ	ظ	x											
ʕ	ع	3											
q	غ	q											
f	ف	f											
q	ق	q											
k	ك	k											
l	ل	l											
m	م	m											
n	ن	n											
h	ه	h											
w	و	w											
y	ي	y											

A.2. Hypocoristics usage with Hybrid Model

sound trans	C1V0C2V0C3V → C1aC2C2uuC3	C1V0C2V0C3V0aan → C1aC2C2uuC3	ʔaC1C2V0C3 → C1aC2C2uuC3	C1V0C2C3V0C4 → C1aC2C3uuC4	ʔiC1V0C2V0C3V0C4 → C1aC2C2uuC3	ʔiC1tV0C2V0C3V → C1aC2C2uuC3	C1V0C2V0C3aʔ → C1aC2C2uuC3	C1V0C2V0jV0C3 → C1aC2C2uuC3	C1V0C2V0wV0 → C1aC2C2uuj	mV0C1V0C2V0C3 → C1aC2C2uuC3	C1V0C2V0→C1aC2C2uuj	C1V0C2V→C1ajjuuC2	C1V0C2→C1ajjuun	C1aC2aaC3iC4 → C1aC2aaC3iC4
basil	L													
bedir	L													
bedriye														
besam	L													
besme	L													
becayir								N						
becar	L													
bicir	L													
2ibtihal						K								
bucre	L													
bu8ene	L													
dane											Q		S	
delal	L													
dawod														
dpme											Q	R		
dpne											Q	R		
jerah	L													
jewahir														T
juwhere														T
xarp											Q			
xu7e											Q			
fatin	L													
fedwe									O					

A.2. Hypocoristics usage with Hybrid Model

fehed	L													
fey													S	
fela7	L													
fe9el	L													
fuz														
qade											Q	R		
hadp											Q			
hale											Q			
hanp											Q			
hedpl	L													
heyfa2							M							
7emed	L													
7amdan		G												
7enan	L													
7esen	L													
7esna2							M							
hind	L													
7sen	L													
hude											Q			
7usam	L													
ye7ye														
yosif	L													
leme											Q			
lemya2							M							
le6pfe	L													
lpn											Q	R		
lujen	L													

A.2. Hypocoristics usage with Hybrid Model

mejd	L													
mehe											Q			
m7muud										P				
mey													S	
mey8a2							M							
men9or										P				
meryem				I										
merwe									O					
mrzog										P				
mece3il										P				
me8ayil								N						
mbarek										P				
m7emed										P				
mp3ad	L													
micari	L													
mc3el										P				
mnpre	L													
msa3ed				I						P				
mufpde	L													
mhened										P				
mune											Q			
muntehe														
mu96efe														
na9ir	L													
nede											Q			
nejpbe	L													
newaf														

A.2. Hypocoristics usage with Hybrid Model

newal														
ne3pme	L													
nihad	L													
nuf														
no7														
nore														
noriye														
ranye	L													
ra2id	L													
rene											Q			
riyam	L													
rpm												R		
rpme											Q			
rule											Q			
salp											Q			
salim	L													
samp											Q			
sare											S		S	
selam	L													
selman		G												
selwe									O					
sema7	L													
semer	L													
sempre	L													
sene											Q			
sena2	L													
suheyle														

A.2. Hypocoristics usage with Hybrid Model

suh4l	L													
sula											Q			
sul6an		G												
sumeye	L													
su3ad	L													
9efa2	L													
9uba7	L													
ce8e											Q			
cadp											Q			
cehed	L													
ceyma2							M							
cey5e	L													
cemeyil								N						
curoq	L													
6eriq	L													
wefa2	L													
wehab	L													
welpd	L													
wespm	L													
wudad	L													
5elid	L													
5elpl	L													
5elfan		G												
zehre	L													
zehre2							M							
zen											Q	R		
zeneb	L													

A.2. Hypocoristics usage with Hybrid Model

2ebrar			H											
2efra7			H											
2e7lam			H											
2e7med			H											
2eymen			H											
2emani	L													
2enfal	L		H											
2enpse	L													
2enwer														
2eryam			H											
2erwe									O					
2esil	L													
2esma2							M							
2e9ayil								N						
2e9ale	L													
2e9pl	L													
2ibtisam						K								
2ibrahp m					J									
2ibtihaj						K								
2pman	L													
2imti8al						K								
2inti9ar						K								
2useme	L													
3ebpr	L													
3ednan		G												
3efaf	L													

A.2. Hypocoristics usage with Hybrid Model

3eyce	L													
3elp											Q			
3ewa6if														T
3ezpz	L													
3pse														
3i9am	L													
3umer	L													
3u8man		G												

A.3 Names and Hypocoristics Formed by D&Z’s Rule

#	name	hypocoristic	#	name	hypocoristics
1	badrija	badduur	71	rijm	rajjum
2	badir	badduur	72	rijma	rajjum
3	ʔibtihaaʤ	bahhuuʤ	73	rijaam	rajjum
4	mbaarak	barruuk	74	raaʔid	rajjuud
5	ʔabraar	barruur	75	ʔarjaam	rajjuum
6	baasil	bassuul	76	raanja	rannuuj
7	basma	bassuum	77	rana	rannuuj
8	basaam	bassuum	78	ʔarwa	rawwuuj
9	ʔbtisaam	bassuum	79	marzuug	razzuug
10	baʃaajir	baʃʃuur	80	suhajla	sahhuul
11	buʃra	baʃʃuur	81	suheil	sahhuul
12	baʃaar	baʃʃuur	82	salwa	salluuj
13	biʃir	baʃʃuur	83	sula	salluuj
14	buθajna	baθθuun	84	saalij	salluuj
15	diima	dajjuum	85	saalim	salluum
16	diina	dajjuun	86	salmaan	salluum
17	dalaal	dalluul	87	salaam	salluum
18	ʤaraah	ʤaruuħ	88	samaaħ	sammuuħ
19	ʤawaahir	ʤawhhuur	89	sumaja	sammuuj
20	ʤohara	ʤawhhuur	90	saami	sammuuj
21	ðˤaari	ðˤarruuj	91	samiira	sammuur
22	ðˤuħa	ðˤuħa	92	samar	sammuur
23	fadwa	fadduuj	93	sanaʔ	sannuuj
24	fahad	fahhuud	94	sana	sannuuj
25	falaaħ	falluuħ	95	saara	sarruur
26	ʔafraaħ	farruuħ	96	msaaʕad	saʕʕuud
27	feisˤal	fasˤsˤuul	97	suʕaad	saʕʕuud
28	faatin	fattuun	98	sˤubaaħ	sˤabbuuħ
29	3aada	3ajjuud	99	sˤafaʔ	sˤaffuuj
30	haadi	hadduuj	100	ʃadi	ʃadduuj
31	huda	hadduuj	101	ʃaða	ʃaððuuj
32	hadiil	hadduul	102	ʃahhad	ʃahhuud
33	hajfaaʔ	hajjuuf	103	ʃajmaʔ	ʃajjuum
34	haala	halluuj	104	ʃajχa	ʃajjuuχ
35	mħamad	ħammuud	105	ʃamajil	ʃammuul
36	ʔaħmad	ħammuud	106	ʃuruuq	ʃarruuq

37	ħamdaan	ħammuud	107	maʃaʕil	ʃaʕʕuul
38	maħmuud	ħammuud	108	miʃʕal	ʃaʕʕuul
39	ħamad	ħammuud	109	tˤariq	tˤarruuq
40	hind	hannuud	110	wudaad	wadduud
41	muhanad	hannuud	111	wafaʔ	waffuuj
42	haani	hannuuj	112	wahaab	wahhuub
43	ħanaan	ħannuun	113	waliid	walluud
44	ħusaam	ħassuum	114	wasiim	wassuum
45	ħsein	ħassuun	115	miiʕaad	waʕʕuud
46	ħasan	ħassuun	116	xalid	xalluud
47	ħasnaaʔ	ħassuun	117	xaliil	xalluul
48	ʔajman	jammuun	118	xalˤfaan	xalˤlˤuuf
49	luʤein	laʤʤuun	119	zahraʔ	zahhuur
50	lamjaaʔ	lammuuj	120	zahra	zahhuur
51	lama	lammuuj	121	zein	zajjuun
52	latˤiifa	latˤtˤuuf	122	zeinab	zannuub
53	maʤd	maʤʤuud	123	ʔiimaan	ʔammuun
54	mufiida	maffuud	124	ʔamaani	ʔamuun
55	maha	mahuuj	125	ʔaniisa	ʔannuus
56	majθaaʔ	majjuuθ	126	ʔusama	ʔassuum
57	muna	mannuuj	127	ʔasmaaʔ	ʔassuum
58	marwa	marruuj	128	ʔasˤiil	ʔasˤsˤuul
59	miʃaari	maʃʃuur	129	ʔasˤaajil	ʔasˤsˤuul
60	maθaajil	maθθuul	130	ʔasˤaala	ʔasˤsˤuul
61	ʔimtiθaal	maθθuul	131	ʕabiir	ʕabbuur
62	nada	nadduuj	132	ʕadnaan	ʕadduun
63	naʤiiba	naʤʤuub	133	ʕafaaf	ʕaffuuf
64	ʔanfaal	naffuul	134	ʕajʃa	ʕajjuuʃ
65	nihaad	nahhuud	135	ʕalij	ʕalluuj
66	nawaaf	najjuuf	136	ʕumar	ʕammuur
67	ʔintisˤaar	nasˤsˤuur	137	ʕisˤaam	ʕasˤsˤuum
68	naasˤir	nasˤsˤuur	138	ʕawaatˤif	ʕatˤtˤuuf
69	mansˤuur	nasˤuur	139	ʕaziiz	ʕazuuz
70	naʕiima	naʕʕuum	140	ʕuθmaan	ʕaθθuum

A.4 Frequencies of data used in correlation tests.

#	name	gender	popular	name_freq	patterns	nicknfreq	Name_LEN
a.01	msa3ed	m	30	38	11	23	6
a.02	fedwe	f	4	3	10	19	5
a.03	bedriye	f	51	100	10	25	7
a.04	besme	f	22	26	6	24	5
a.05	becayir	f	67	218	9	29	7
a.06	besam	m	7	6	7	21	5
a.07	bu84ne	f	4	3	5	13	6
a.08	wefa2	f	26	33	7	16	5
a.09	newaf	m	44	85	8	22	5
a.10	me8ayil	f	9	8	9	18	7
a.11	cey5e	f	75	318	9	31	5
a.12	curoq	f	43	82	7	19	5
a.13	hind	f	41	75	5	20	4
a.14	3eliy	m	77	376	8	42	5
a.15	ceyma2	f	52	104	5	23	6
a.16	2esil	f	40	74	6	25	5
a.17	riym	f	83	466	9	27	4
a.18	men9or	m	24	28	5	12	6
a.19	diyme	f	42	81	7	22	5
a.20	qade	f	16	16	6	15	4
a.21	xariy	m	36	65	3	20	5
a.22	xu7e	f	46	90	5	17	4
a.23	nejpbe	f	4	3	4	17	6
a.24	mnpre	f	81	442	7	28	5
a.25	2useme	m	13	13	3	13	6
a.26	muntehe	f	3	2	4	5	7
a.27	mp3ad	f	6	5	4	7	5
a.28	bedir	m	64	184	10	35	5
a.29	2ebrar	f	69	229	7	24	6
a.30	sul6an	m	47	93	6	20	6
a.31	zehre2	f	29	36	10	31	6
a.32	fey	f	45	86	9	19	3
a.33	2nti9ar	f	21	24	4	10	7
a.34	3ewa6if	f	19	22	7	18	7
a.35	3ezpz	m	65	191	6	29	5
b.01	na9ir	m	69	229	6	43	5
b.02	merwe	f	18	20	9	21	5

b.03	sumeye	f	17	17	9	24	6
b.04	bucre	f	10	9	7	24	5
b.05	cemeyil	f	30	38	8	27	7
b.06	becar	m	2	1	9	25	5
b.07	suheyle	f	4	3	6	17	7
b.08	sena2	f	12	12	6	19	5
b.09	newal	f	38	69	7	35	5
b.10	mece3il	f	49	95	7	16	7
b.11	3eyce	f	80	411	13	35	5
b.12	7usam	m	3	2	7	19	5
b.13	mejd	m	3	2	7	24	4
b.14	2enwer	m	7	6	5	7	6
b.15	mey8a2	f	6	5	3	11	6
b.16	2e9pl	m	2	1	8	16	5
b.17	zeyn	f	62	154	8	21	4
b.18	merzog	m	17	17	9	13	6
b.19	diyne	f	15	15	10	20	5
b.20	hale	f	6	5	6	11	4
b.21	hadiy	m	15	15	7	21	5
b.22	nede	f	37	66	9	26	4
b.23	ne3pme	f	8	7	8	19	6
b.24	mufpde	f	1	0	9	9	6
b.25	2enpse	f	2	1	5	15	6
b.26	mu96efe	m	10	9	8	14	7
b.27	2pman	f	70	253	8	23	5
b.28	fehed	m	78	384	7	33	5
b.29	2efra7	f	50	97	7	24	6
b.30	5elfan	m	1	0	4	14	6
b.31	2esma2	f	76	328	7	21	6
b.32	mey	f	46	90	11	23	3
b.33	2btisam	f	29	36	9	19	7
b.34	jewahir	f	35	62	6	14	7
b.35	5elpl	m	8	7	6	21	5
c.01	salim	m	48	94	6	28	5
c.02	2erwe	f	28	35	8	20	5
c.03	ranye	f	5	4	9	28	5
c.04	zehre	f	13	13	9	33	5
c.05	2e9ayil	f	24	28	8	20	7
c.06	delal	f	84	536	8	32	5
c.07	7s4n	m	60	151	10	37	4
c.08	9efa2	f	5	4	9	21	5
c.09	mic3el	m	55	111	10	25	6
c.10	nihad	f	1	0	4	24	5

c.11	2eymen	m	1	0	7	14	6
c.12	heyfa2	f	17	17	8	17	6
c.13	nuf	f	73	295	12	26	3
c.14	3pse	m	33	57	10	25	4
c.15	sare	f	87	1215	8	36	4
c.16	samiy	m	14	14	8	24	5
c.17	rene	f	8	7	8	25	4
c.18	le6pfe	f	63	173	7	27	6
c.19	2e9ale	f	2	1	8	21	6
c.20	7esen	m	39	73	8	32	5
c.21	2e7lam	f	22	26	6	18	6
c.22	selman	m	37	66	9	28	6
c.23	7esna2	f	9	8	12	20	6
c.24	hedpl	f	56	115	8	22	5
c.25	m7emed	m	86	1093	11	35	6
c.26	z4neb	f	71	258	9	31	5
c.27	dawod	m	7	6	10	17	5
c.28	micariy	m	53	106	8	22	7
c.29	wudad	f	14	14	9	21	5
c.30	2brahpm	m	50	97	6	28	7
c.31	fatin	f	14	14	8	24	5
c.32	ye7ye	m	4	3	10	23	5
c.33	sema7	f	2	1	8	22	5
c.34	su3ad	f	25	29	8	19	5
c.35	mune	f	59	150	11	30	4
d.01	6eriq	m	11	11	8	32	5
d.02	selwe	f	28	35	9	26	5
d.03	noriye	f	10	9	12	33	6
d.04	7enan	f	65	191	10	41	5
d.05	suh4l	m	1	0	8	26	5
d.06	jera7	m	31	39	10	27	5
d.07	2e7med	m	82	443	12	49	6
d.08	fuz	f	10	9	7	27	3
d.09	riyme	f	8	7	12	37	5
d.10	mehe	f	72	267	7	30	4
d.11	sempre	f	9	8	14	33	6
d.12	3umer	m	58	146	7	28	5
d.13	2eryam	f	1	0	11	22	6
d.14	3u8man	m	27	34	8	41	6
d.15	lemya2	f	18	20	10	26	6
d.16	welpd	m	32	53	9	31	5
d.17	mbarek	m	57	117	6	27	6
d.18	f49el	m	66	192	9	32	5

d.19	3ebpr	f	61	153	6	32	5
d.20	yosif	m	74	298	4	34	5
d.21	ra2id	m	5	4	3	20	5
d.22	2emaniy	f	54	108	8	25	7
d.23	hude	f	34	60	5	18	4
d.24	luj4n	f	20	23	7	26	5
d.25	juhere	f	4	3	5	20	6
d.26	rule	f	1	0	6	14	4
d.27	5elid	m	79	392	9	35	5
d.28	mhened	m	7	6	3	10	6
d.29	riyam	f	1	0	6	19	5
d.30	fela7	m	23	27	5	24	5
d.31	3ednan	m	10	9	8	19	6
d.32	2enfal	f	68	225	10	28	6
d.33	3efaf	f	12	12	7	21	5
d.34	nore	f	85	854	10	39	4
d.35	9uba7	m	2	1	6	26	5

A.5 Example of a data file default hypocoristic

```
A 0Cm==CsVaC3VeCd==== msa3ed
A 1CfVeCd==CwVe====== fedwe
A 1CfVeCd==CwVe====== fedwe
A 1CfVeCd==CwVe====== fedwe
A 1CfVeCd==CwVe====== fedwe
C 1CfVeCd==CwVe====== fedwe
C 1CfVeCd==CwVe====== fedwe
J 1CfVeCd==CwVe====== fedwe
J 1CfVeCd==CwVe====== fedwe
A 1CbVeCd==CrViCyVe== bedriye
A 1CbVeCd==CrViCyVe== bedriye
C 1CbVeCd==CrViCyVe== bedriye
C 1CbVeCd==CrViCyVe== bedriye
B 1CbVeCcVaCyViCr==== becayir
F 1CbVeCcVaCyViCr==== becayir
F 1CbVeCcVaCyViCr==== becayir
B 1CwVeCfVaC2======== wefa2
B 1CwVeCfVaC2======== wefa2
B 1CwVeCfVaC2======== wefa2
B 1CwVeCfVaC2======== wefa2
B 1CwVeCfVaC2======== wefa2
A 0CnVeCwVaCf======== newaf
A 0CnVeCwVaCf======== newaf
A 0CnVeCwVaCf======== newaf
A 0CnVeCwVaCf======== newaf
A 0CnVeCwVaCf======== newaf
A 0CnVeCwVaCf======== newaf
A 1CmVeC8VaCyViCl==== me8ayil
A 1CmVeC8VaCyViCl==== me8ayil
C 1CmVeC8VaCyViCl==== me8ayil
I 1CmVeC8VaCyViCl==== me8ayil
I 1CmVeC8VaCyViCl==== me8ayil
A 1CcVeCy==C5Ve====== cey5e
A 1CcVeCy==C5Ve====== cey5e
A 1CcVuCrVoCq======== curoq
A 1CcVuCrVoCq======== curoq
A 1CcVuCrVoCq======== curoq
A 1CcVuCrVoCq======== curoq
B 1ChViCn==Cd======== hind
B 1ChViCn==Cd======== hind
A 1CcVeCy==CmVaC2==== ceyma2
A 1CcVeCy==CmVaC2==== ceyma2
A 1CcVeCy==CmVaC2==== ceyma2
A 1CcVeCy==CmVaC2==== ceyma2
B 1C2VeCsViCl======== 2esil
B 1C2VeCsViCl======== 2esil
A 1CrViCy==Cm======== riym
A 1CrViCy==Cm======== riym
A 1CrViCy==Cm======== riym
C 1CrViCy==Cm======== riym
C 1CrViCy==Cm======== riym
C 1CrViCy==Cm======== riym
D 1CrViCy==Cm======== riym
D 1CrViCy==Cm======== riym
H 0CmVeCn==C9VoCr==== men9or
A 1CdViCy==CmVe====== diyme
```

A.6 Example of Test file used in Affix –o and CVCV Hypocoristics

```
UNK 0C3ViC9VaCm======== 3i9am

UNK 0ChVaCnViCy======== haniy

UNK 1CsVeCnVe========== sene

UNK 0CcVaCdVi========== cadi

UNK 1CsVeClVaCm======== selam

UNK 1CqVaCdVe========== qada

UNK 0CxVaCrViCy======== xari

UNK 1CxVuC7Ve========== xu7a

UNK 0CsVaClViCm======== salim

UNK 1C2VeC7==ClVaCm==== 2a7lam

UNK 1ChVeCdVpCl======== hedpl

UNK 0CwVeCsVpCm======== wespm

UNK 1CbVeCs==CmVe====== besme

UNK 0C3VeClViCy======== 3eliy

UNK 1CmVeCy==C8VaC2==== mey8a2
```

A.7 Example of result file for a test of 2 names 3ia9am and haniy with default hypocoristics

```
///---------------------///

3i9am:

Gang summary (debug):
+------------+--------+-----------+-------+----------------------------
-----------+--------------+
| Percentage | Score  | Num Items | Class |
| Item Comment |
| Context    |        |           |       | 0 C 3 V i C 9 V a C m
|              |
+------------+--------+-----------+-------+----------------------------
-----------+--------------+
***********************************************************************
***************************
| 45.400     | 913408 |           |       | 0 C * V * C 9 V * C * * * *
* * * * * |              |
+------------+--------+-----------+-------+----------------------------
-----------+--------------+
| 11.350     | 114176 | 2         | A     |
|              |
|            |        |           |       | 0 C n V a C 9 V i C r
| na9ir        |
|            |        |           |       | 0 C n V a C 9 V i C r
| na9ir        |
| 11.350     | 114176 | 2         | B     |
|              |
|            |        |           |       | 0 C 2 V e C 9 V p C l
| 2e9pl        |
|            |        |           |       | 0 C 2 V e C 9 V p C l
| 2e9pl        |
| 11.350     | 114176 | 2         | H     |
|              |
|            |        |           |       | 0 C 2 V e C 9 V p C l
| 2e9pl        |
|            |        |           |       | 0 C 2 V e C 9 V p C l
| 2e9pl        |
| 11.350     | 114176 | 2         | K     |
|              |
|            |        |           |       | 0 C 2 V e C 9 V p C l
| 2e9pl        |
|            |        |           |       | 0 C 2 V e C 9 V p C l
| 2e9pl        |
***********************************************************************
***************************
| 39.089     | 786432 |           |       | * C * V i C * V a C m * * *
* * * * * |              |
+------------+--------+-----------+-------+----------------------------
-----------+--------------+
| 14.658     |  98304 | 3         | A     |
|              |
|            |        |           |       | 1 C r V i C y V a C m
| riyam        |
|            |        |          |       | 1 C r V i C y V a C m
| riyam        |
|            |        |          |       | 1 C r V i C y V a C m
| riyam        |
|  4.886     |  98304 | 1        | C     |
|              |
|            |        |          |       | 1 C r V i C y V a C m
| riyam        |
| 14.658     |  98304 | 3        | G     |
|              |
|            |        |          |       | 1 C r V i C y V a C m
| riyam        |
|            |        |          |       | 1 C r V i C y V a C m
| riyam        |
|            |        |          |       | 1 C r V i C y V a C m
| riyam        |
|  4.886     |  98304 | 1        | I     |
|              |
|            |        |          |       | 1 C r V i C y V a C m
| riyam        |
**********************************************************************
**************************
|  6.515     | 131072 |          |       | * C 3 V * C * V a C * * * *
* * * * * |             |
+------------+--------+----------+-------+----------------------------
-----------+-------------+
|  4.886     |  32768 | 3        | B     |
|              |
|            |        |          |       | 1 C 3 V e C f V a C f
| 3efaf        |
|            |        |          |       | 1 C 3 V e C f V a C f
| 3efaf        |
|            |        |          |       | 1 C 3 V e C f V a C f
| 3efaf        |
|  1.629     |  32768 | 1        | F     |
|              |
|            |        |          |       | 1 C 3 V e C f V a C f
| 3efaf        |
**********************************************************************
**************************
|  6.311     | 126976 |          |       | 0 C 3 V * C * V * C * * * *
* * * * * |             |
+------------+--------+----------+-------+----------------------------
-----------+-------------+
|  3.156     |  31744 | 2        | B     |
|              |
|            |        |          |       | 0 C 3 V e C z V p C z
| 3ezpz        |
|            |        |          |       | 0 C 3 V e C z V p C z
| 3ezpz        |
|  3.156     |  31744 | 2        | K     |
|              |
|            |        |          |       | 0 C 3 V e C z V p C z
| 3ezpz        |
|            |        |          |       | 0 C 3 V e C z V p C z
| 3ezpz        |
**********************************************************************
**************************
```

```
|  1.578     |  31744 |           |       | 0 C * V i C * V * C * * * *
* * * * * |              |
+------------+--------+-----------+-------+---------------------------
-----------+--------------+
|  1.578     |  15872 | 2         | A     |
|              |
|            |        |           |       | 0 C b V i C c V i C r
| bicir        |
|            |        |           |       | 0 C b V i C c V i C r
| bicir        |
***********************************************************************
***************************
|  0.802     |  16128 |           |       | 0 C * V * C * V a C m * * *
* * * * * |              |
+------------+--------+-----------+-------+---------------------------
-----------+--------------+
|  0.802     |  16128 | 1         | A     |
|              |
|            |        |           |       | 0 C 7 V u C s V a C m
| 7usam        |
***********************************************************************
***************************
|  0.305     |   6144 |           |       | * C * V * C 9 V a C * * * *
* * * * * |              |
+------------+--------+-----------+-------+---------------------------
-----------+--------------+
|  0.153     |   3072 | 1         | B     |
|              |
|            |        |           |       | 1 C 2 V e C 9 V a C l V e
| 2e9ale       |
|  0.153     |   3072 | 1         | H     |
|              |
|            |        |           |       | 1 C 2 V e C 9 V a C l V e
| 2e9ale       |
+------------+--------+-----------+-------+---------------------------
-----------+--------------+

Winners:
A
Scores normalized:
$VAR1 = {
          'F' => '0.0162870594223184',
          'A' => '0.28387835602494',
          'H' => '0.115027357170123',
          'K' => '0.145056622980023',
          'C' => '0.0488611782669551',
          'G' => '0.146583534800865',
          'B' => '0.19544471306782',
          'I' => '0.0488611782669551'
        };
Statistical Summary
+-------+---------+------------+
| Class | Score   | Percentage |
+-------+---------+------------+
| A     |  571136 | 28.388     |
| B     |  393216 | 19.544     |
| C     |   98304 |  4.886     |
| F     |   32768 |  1.629     |
| G     |  294912 | 14.658     |
| H     |  231424 | 11.503     |
| I     |   98304 |  4.886     |
| K     |  291840 | 14.506     |
+-------+---------+------------+
| Total | 2011904 |            |
+-------+---------+------------+
Expected class unknown

Analogical summary:
Analogical Set
Total Frequency = 2011904
+-------+--------+--------+------------+
| Class | Item   | Score  | Percentage |
+-------+--------+--------+------------+
| B     | 3ezpz  |  31744 | 1.578      |
| B     | 3ezpz  |  31744 | 1.578      |
| K     | 3ezpz  |  31744 | 1.578      |
| K     | 3ezpz  |  31744 | 1.578      |
| A     | na9ir  | 114176 | 5.675      |
| A     | na9ir  | 114176 | 5.675      |
| A     | 7usam  |  16128 | 0.802      |
| B     | 2e9pl  | 114176 | 5.675      |
| B     | 2e9pl  | 114176 | 5.675      |
| H     | 2e9pl  | 114176 | 5.675      |
| H     | 2e9pl  | 114176 | 5.675      |
| K     | 2e9pl  | 114176 | 5.675      |
| K     | 2e9pl  | 114176 | 5.675      |
| B     | 2e9ale |   3072 | 0.153      |
| H     | 2e9ale |   3072 | 0.153      |
| A     | riyam  |  98304 | 4.886      |
| A     | riyam  |  98304 | 4.886      |
| A     | riyam  |  98304 | 4.886      |
| C     | riyam  |  98304 | 4.886      |
| G     | riyam  |  98304 | 4.886      |
| G     | riyam  |  98304 | 4.886      |
| G     | riyam  |  98304 | 4.886      |
| I     | riyam  |  98304 | 4.886      |
| B     | 3efaf  |  32768 | 1.629      |
| B     | 3efaf  |  32768 | 1.629      |
| B     | 3efaf  |  32768 | 1.629      |
| F     | 3efaf  |  32768 | 1.629      |
| A     | bicir  |  15872 | 0.789      |
| A     | bicir  |  15872 | 0.789      |
+-------+--------+--------+------------+

///---------------------///

haniy:

Gang summary (debug):
+------------+-------+----------+-------+--------------------------
-----------+-------------+
| Percentage | Score  | Num Items | Class |
| Item Comment |
| Context    |        |           |       | 0 C h V a C n V i C y
|              |
+------------+--------+----------+-------+-----------------------------
-----------+-------------+
************************************************************************
***************************
| 45.433     | 501760 |          |       | * C * V a C n * * C y * * *
* * * * * |               |
+------------+--------+----------+-------+-----------------------------
-----------+-------------+
| 19.471     |  35840 | 6        | A     |
|              |
|            |        |          |       | 1 C r V a C n     C y V e
| ranye        |
|            |        |          |       | 1 C r V a C n     C y V e
| ranye        |
|            |        |          |       | 1 C r V a C n     C y V e
| ranye        |
|            |        |          |       | 1 C r V a C n     C y V e
| ranye        |
|            |        |          |       | 1 C r V a C n     C y V e
| ranye        |
|            |        |          |       | 1 C r V a C n     C y V e
| ranye        |
|  6.490     |  35840 | 2        | C     |
|              |
|            |        |          |       | 1 C r V a C n     C y V e
| ranye        |
|            |        |          |       | 1 C r V a C n     C y V e
| ranye        |
| 19.471     |  35840 | 6        | G     |
|              |
|            |        |          |       | 1 C r V a C n     C y V e
| ranye        |
|            |        |          |       | 1 C r V a C n     C y V e
| ranye        |
|            |        |          |       | 1 C r V a C n     C y V e
| ranye        |
|            |        |          |       | 1 C r V a C n     C y V e
| ranye        |
|            |        |          |       | 1 C r V a C n     C y V e
| ranye        |
|            |        |          |       | 1 C r V a C n     C y V e
| ranye        |
************************************************************************
***************************
| 31.803     | 351232 |          |       | * C * V * C n V * C * * * *
* * * * * |               |
+------------+--------+----------+-------+-----------------------------
-----------+---------------+
| 15.902     |  25088 | 7        | A     |
|              |
|            |        |          |       | 1 C z V 4 C n V e C b
| z4neb        |
|            |        |          |       | 1 C z V 4 C n V e C b
| z4neb        |
|            |        |          |       | 1 C z V 4 C n V e C b
| z4neb        |
|            |        |          |       | 1 C z V 4 C n V e C b
| z4neb        |
|            |        |           |       | 1 C z V 4 C n V e C b
| z4neb        |
|            |        |           |       | 1 C s V e C n V a C 2
| sena2        |
|            |        |           |       | 1 C s V e C n V a C 2
| sena2        |
|  2.272     |  25088 | 1         | B     |
|              |
|            |        |           |       | 1 C 7 V e C n V a C n
| 7enan        |
|  2.272     |  25088 | 1         | C     |
|              |
|            |        |           |       | 1 C z V 4 C n V e C b
| z4neb        |
| 11.358     |  25088 | 5         | K     |
|              |
|            |        |           |       | 1 C z V 4 C n V e C b
| z4neb        |
|            |        |           |       | 1 C z V 4 C n V e C b
| z4neb        |
|            |        |           |       | 1 C z V 4 C n V e C b
| z4neb        |
|            |        |           |       | 1 C z V 4 C n V e C b
| z4neb        |
|            |        |           |       | 1 C z V 4 C n V e C b
| z4neb        |
***********************************************************************
***************************
|  6.073     |  67072 |           |       | 0 C * V a C * V i C y * * *
* * * * * |              |
+------------+--------+-----------+-------+----------------------------
-----------+--------------+
|  6.073     |  67072 | 1         | A     |
|              |
|            |        |           |       | 0 C s V a C m V i C y
| samiy        |
***********************************************************************
***************************
|  5.192     |  57344 |           |       | * C * V a C n V * * * * * *
* * * * * |              |
+------------+--------+-----------+-------+----------------------------
-----------+--------------+
|  5.192     |  14336 | 4         | A     |
|              |
|            |        |           |       | 1 C d V a C n V e
| dane         |
|            |        |           |       | 1 C d V a C n V e
| dane         |
|            |        |           |       | 1 C d V a C n V e
| dane         |
|            |        |           |       | 1 C d V a C n V e
| dane         |
***********************************************************************
***************************
|  3.616     |  39936 |           |       | * C h V a C * V * * * * * *
* * * * * |              |
+------------+--------+-----------+-------+----------------------------
-----------+--------------+
|  3.616     |  19968 | 2         | B     |
|              |
|            |        |           |       | 1 C h V a C l V e
| hale         |
|            |        |           |       | 1 C h V a C l V e
| hale         |
**********************************************************************
**************************
|  3.570     |  39424 |           |       | * C h V * C n * * C * * * *
* * * * * |              |
+------------+--------+-----------+-------+---------------------------
-----------+--------------+
|  3.570     |  19712 | 2         | B     |
|              |
|            |        |           |       | 1 C h V i C n    C d
| hind         |
|            |        |           |       | 1 C h V i C n    C d
| hind         |
**********************************************************************
**************************
|  2.782     |  30720 |           |       | 0 C * V * C * * * C y * * *
* * * * * |              |
+------------+--------+-----------+-------+---------------------------
-----------+--------------+
|  2.782     |   6144 | 5         | A     |
|              |
|            |        |           |       | 0 C y V e C 7    C y V e
| ye7ye        |
|            |        |           |       | 0 C y V e C 7    C y V e
| ye7ye        |
|            |        |           |       | 0 C y V e C 7    C y V e
| ye7ye        |
|            |        |           |       | 0 C y V e C 7    C y V e
| ye7ye        |
|            |        |           |       | 0 C y V e C 7    C y V e
| ye7ye        |
**********************************************************************
**************************
|  0.834     |   9216 |           |       | 0 C * V * C n * * C * * * *
* * * * * |              |
+------------+--------+-----------+-------+---------------------------
-----------+--------------+
|  0.278     |   1536 | 2         | A     |
|              |
|            |        |           |       | 0 C m V e C n    C 9 V o C
r         | men9or      |
|            |        |           |       | 0 C m V e C n    C 9 V o C
r         | men9or      |
|  0.417     |   1536 | 3         | B     |
|              |
|            |        |           |       | 0 C 2 V e C n    C w V e C
r         | 2enwer      |
|            |        |           |       | 0 C 2 V e C n    C w V e C
r         | 2enwer      |
|            |        |           |       | 0 C 2 V e C n    C w V e C
r         | 2enwer      |
|  0.139     |   1536 | 1         | H     |
|              |
|           |        |           |       | 0 C m V e C n     C 9 V o C
r         | men9or       |
************************************************************************
***************************
|  0.695     |   7680 |           |       | * C * V * C * V * C y * * *
* * * * * |              |
+-----------+-------+----------+-------+--------------------------
-----------+-------------+
|  0.695     |   1920 | 4         | A     |
|              |
|           |        |           |       | 1 C s V u C h V e C y     C
l V e     | suheyle      |
|           |        |           |       | 1 C s V u C h V e C y     C
l V e     | suheyle      |
|           |        |           |       | 1 C s V u C h V e C y     C
l V e     | suheyle      |
|           |        |           |       | 1 C s V u C h V e C y     C
l V e     | suheyle      |
+-----------+-------+----------+-------+--------------------------
-----------+-------------+

Winners:
A
Scores normalized:
$VAR1 = {
          'C' => '0.0876216968011127',
          'G' => '0.19471488178025',
          'B' => '0.0987482614742698',
          'A' => '0.503940658321743',
          'H' => '0.00139082058414465',
          'K' => '0.113583681038479'
        };
Statistical Summary
+-------+---------+------------+
| Class | Score   | Percentage |
+-------+---------+------------+
| A     |  556544 | 50.394     |
| B     |  109056 |  9.875     |
| C     |   96768 |  8.762     |
| G     |  215040 | 19.471     |
| H     |    1536 |  0.139     |
| K     |  125440 | 11.358     |
+-------+---------+------------+
| Total | 1104384 |            |
+-------+---------+------------+
Expected class unknown

Analogical summary:
Analogical Set
Total Frequency = 1104384
+-------+---------+-------+------------+
| Class | Item    | Score | Percentage |
+-------+---------+-------+------------+
| B     | hind    | 19712 | 1.785      |
| B     | hind    | 19712 | 1.785      |
| H     | men9or  |  1536 | 0.139      |
| A     | suheyle |  1920 | 0.174      |
| A     | suheyle |  1920 | 0.174      |
| A     | suheyle |  1920 | 0.174      |
| A     | suheyle |  1920 | 0.174      |
| A     | sena2   | 25088 | 2.272      |
| A     | sena2   | 25088 | 2.272      |
| B     | 2enwer  |  1536 | 0.139      |
| B     | 2enwer  |  1536 | 0.139      |
| B     | 2enwer  |  1536 | 0.139      |
| B     | hale    | 19968 | 1.808      |
| B     | hale    | 19968 | 1.808      |
| A     | ranye   | 35840 | 3.245      |
| A     | ranye   | 35840 | 3.245      |
| A     | ranye   | 35840 | 3.245      |
| A     | ranye   | 35840 | 3.245      |
| A     | ranye   | 35840 | 3.245      |
| A     | ranye   | 35840 | 3.245      |
| C     | ranye   | 35840 | 3.245      |
| C     | ranye   | 35840 | 3.245      |
| G     | ranye   | 35840 | 3.245      |
| G     | ranye   | 35840 | 3.245      |
| G     | ranye   | 35840 | 3.245      |
| G     | ranye   | 35840 | 3.245      |
| G     | ranye   | 35840 | 3.245      |
| G     | ranye   | 35840 | 3.245      |
| A     | samiy   | 67072 | 6.073      |
| A     | z4neb   | 25088 | 2.272      |
| A     | z4neb   | 25088 | 2.272      |
| A     | z4neb   | 25088 | 2.272      |
| A     | z4neb   | 25088 | 2.272      |
| A     | z4neb   | 25088 | 2.272      |
| C     | z4neb   | 25088 | 2.272      |
| K     | z4neb   | 25088 | 2.272      |
| K     | z4neb   | 25088 | 2.272      |
| K     | z4neb   | 25088 | 2.272      |
| K     | z4neb   | 25088 | 2.272      |
| K     | z4neb   | 25088 | 2.272      |
| A     | ye7ye   |  6144 | 0.556      |
| A     | ye7ye   |  6144 | 0.556      |
| A     | ye7ye   |  6144 | 0.556      |
| A     | ye7ye   |  6144 | 0.556      |
| A     | ye7ye   |  6144 | 0.556      |
| B     | 7enan   | 25088 | 2.272      |
| A     | dane    | 14336 | 1.298      |
| A     | dane    | 14336 | 1.298      |
| A     | dane    | 14336 | 1.298      |
| A     | dane    | 14336 | 1.298      |
| A     | men9or  |  1536 | 0.139      |
| A     | men9or  |  1536 | 0.139      |
+-------+---------+-------+------------+

///---------------------///
```

References

762: Analogies. (2017). Retrieved from https://www.explainxkcd.com/wiki/index.php/762:_Analogies

Adams, M. (2009). Power, politeness, and the pragmatics of nicknames. *Names, 57*(2), 81-91.

Allan, K. (1986). *Linguistic meaning*. London: Routledge & Kegan Paul.

Anderson, S. R. A short history of morphological theory. In: The oxford handbook of morphological theory. Oxford: Oxford University Press.

Anttila, R. (2008). Analogy: The Warp and Woof of Cognition. In *The Handbook of Historical Linguistics* (pp. 423-440): Blackwell Publishing Ltd.

Anttila, R., & Brewer, W. A. (1977). *Analogy : a basic bibliography*. Amsterdam: Benjamins.

Arndt-Lappe, S. (2015). Word-formation and analogy. *Müller et al.(Hgg.). Bd, 2,* 822-841.

Aronoff, M. (1976). *Word formation in generative grammar*. Cambridge, Mass.: MIT Press.

Aronoff, M., & Fudeman, K. A. (2011). *What is morphology?* (2nd ed. ed.). Oxford: Wiley-Blackwell.

Aronoff, M., & Lindsay, M. (2016). *Competition and the lexicon.* Paper presented at the Livelli di Analisi e fenomeni di interfaccia. Atti del XLVII congresso internazionale della società di linguistica Italiana.

Bagemihl, B. (1995). Language games and related areas. In e handbook of phonological theory, ed. John Goldsmith, 697–712. In: Cambridge: Blackwell.

Bat-El, O. (2005). The emergence of the trochaic foot in Hebrew hypocoristics. *Phonology, 22*(2), 115-143.

Bat-El, O. (2011). Semitic templates. *Blackwell companion to phonology*, 2586-2608.

Bauer, L. (2001). *Morphological productivity*. Cambridge: Cambridge University Press.

Bauer, L. (2005). Productivity: Theories. In P. Štekauer & R. Lieber (Eds.), *Handbook of word-formation* (pp. 315-334). Dordrecht: Springer Netherlands.

Béland, J. F. P. A. I. R. (2009). *Form and meaning in Arabic names*. Paper presented at the Linguistics in the Gulf II, Qatar.

Blevins, J. P., & Blevins, J. (2009). *Analogy in grammar : form and acquisition*. Oxford: Oxford University Press.

Bonin, P. (2003). *Mental lexicon : some words to talk about words*. Hauppauge, N.Y.: Nova Science Publishers.

Brédart, S. (1993). Retrieval failures in face naming. *Memory, 1*(4), 351-366.

Brennen, T. (1993). The difficulty with recalling people's names: The plausible phonology hypothesis. *Memory, 1*(4), 409-431.

Brylla, E. (2016). Bynames and nicknames. In *The Oxford Handbook of Names and Naming*.

Bybee, J. L. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, J. L. (2007). *Frequency of use and the organization of language*. New York ; Oxford: Oxford University Press.

Bybee, J. L. (2010). *Language, usage and cognition*. Cambridge: Cambridge University Press.

Carstairs-McCarthy, A. (2002). *Current morphology*: Routledge.

Chandler, S. (2009). Exemplar-based models. *Experimental quantitative linguistics*, 100-158.

Chomsky, N. (1965). *Aspects of the theory of syntax. (Second printing.)*. Cambridge, Mass: M.I.T. Press.

Chomsky, N. (1986). *Knowledge of language : its nature, origins, and use*. New York: Praeger.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*: New York, London: Harper & Row.

Cohn, A. C. (2004). *Truncation in Indonesian: Evidence for violable minimal words and AnchorRight.* Paper presented at the PROCEEDINGS-NELS.

Daelemans, W., & Bosch, A. v. d. (2005). *Memory-based language processing*. Cambridge: Cambridge University Press.

Danks, W. (2011). *The Arabic verb: form and meaning in the vowel-lengthening patterns* (Vol. 63): John Benjamins Publishing.

Davis, S., & Tsujimura, N. (2014). Non-concatenative Derivation: Other Processes. *The Oxford Handbook of Derivational Morphology, Oxford University Press, chapter*.

Davis, S., & Zawaydeh, B. (1999). A descriptive analysis of hypocoristics in Colloquial Arabic. *Language and Linguistics, 3*, 83-98.

Dressler, W. U. (2007). *The acquisition of diminutives: A cross-linguistic perspective* (Vol. 43): John Benjamins Publishing.

Dressler, W. U., & Barbaresi, L. M. (1994). *Morphopragmatics: Diminutives and intensifiers in Italian, German, and other languages* (Vol. 76): Walter de Gruyter.

Eddington, D. (2000). Analogy and the dual-route model of morphology. *Lingua, 110*(4), 281-298.

Eddington, D. (2004). *Spanish phonology and morphology : experimental and quantitative perspectives*. Amsterdam: John Benjamin Publishing Company.

Eddington, D. (2009). Linguistic Processing is Exemplar-Based. *Studies in Hispanic and Lusophone Linguistics, 2*(2), 419-434.

Feller, S. (2010). *Lexical meaning in dialogic language use*. Amsterdam ; Philadelphia: John Benjamins Pub. Co.

Geeraerts, D. (2006). *Cognitive linguistics : basic readings*. Berlin ; New York: Mouton de Gruyter.

Guy, G. R. (2011). The SAGE Handbook of Sociolinguistics. In. London: SAGE Publications Ltd. Retrieved from http://sk.sagepub.com/reference/hdbk_sociolinguistics. doi:10.4135/9781446200957

Guy, G. R. (2014). Linking usage and grammar: Generative phonology, exemplar theory, and variable rules. *Lingua, 142*(Supplement C), 57-65. doi:https://doi.org/10.1016/j.lingua.2012.07.007

Hajdú, M. (2003). The History of Onomastics. *Osiris Kiadó, Budapest*.

Heine, B., Narrog, H., Bybee, J. L., & Beckner, C. (2009). Usage-Based Theory. In.

Holland, T. J. (1990). The Many Faces of Nicknames. *Names, 38*(4), 255-272. doi:10.1179/nam.1990.38.4.255

Idrissi, A., Prunet, J.-F., & Béland, R. (2008). On the mental representation of Arabic roots. *Linguistic Inquiry, 39*(2), 221-259.

Ion, T. (n.d.) CHILDREN AS CREATORS OF NAMES. Retreived from http://cis01.central.ucv.ro/revista_scol/site_ro/2008/ion_toma.pdf

Itkonen, E. (2005). *Analogy as structure and process : approaches in linguistics, cognitive psychology and philosophy of science*. Amsterdam: John Benjamins.

Jackendoff, R. (2002). *Foundations of language : brain, meaning, grammar, evolution*. Oxford: Oxford University Press.

Jensen, J. T. (1990). *Morphology : word structure in generative grammar*. Amsterdam: John Benjamins.

Joan, L. B. (2007). From Usage to Grammar: The Mind's Response to Repetition. *Language, 82*(4), 711-733.

Katamba, F., & Stonham, J. T. (2006). *Morphology* (2nd ed. ed.). Basingstoke: Palgrave Macmillan.

Kennedy, R., & Zamuner, T. (2006). Nicknames and the lexicon of sports. *American Speech, 81*(4), 387-422.

Kolers, P. A., & Roediger, H. L. (1984). Procedures of mind. *Journal of Verbal Learning and Verbal Behavior, 23*(4), 425-449.

Kurt, W. (2017). Kullback-Leibler Divergence Explained. Retrieved from https://www.countbayesie.com/blog/2017/5/9/kullback-leibler-divergence-explained

Langendonck, W. v. (2007). *Theory and typology of proper names*. Berlin ; New York: Mouton de Gruyter.

Lappe, S. (2007). English Prosodic Morphology: Morphology. In *English Prosodic Morphology* (pp. 3-30). Dordrecht: Springer Netherlands.

Leibring, K. (2016). Given Names in European Naming Systems. In *The Oxford Handbook of Names and Naming*.(pp.199-213).

Lieber, R. (2014). Theoretical approaches to derivation.

Lieber, R. (2015). 7. Word-formation in generative grammar. In *Word-FormationAn International Handbook of the Languages of Europe*.

Lieber, R., Štekauer, P., Aronoff, M., & Lindsay, M. Productivity, Blocking, and Lexicalization. In.

Lipski, J. M. (1995). Spanish hypocoristics: towards a unified prosodic analysis. *Hispanic Linguistics, 6*, 387-434.

Littlemore, J. e., & Taylor, J. R. D. e. (2014). *The Bloomsbury Companion to Cognitive Linguistics*.

Lowe, A. (2004). Two Syllables Are Better Than One: A Prosodic Template for Bengali Hypocoristics. *Working Papers of the Linguistics Circle, 18*(1), 74-83.

Mattiello, E. (2017). *Analogy in Word-formation, A Study of English Neologisms and Occasionalisms*.

Mattiello, E. a. (2013). *Extra-grammatical morphology in English : abbreviations, blends, reduplicatives, and related phenomena*.

McAndrew, A. (1992). Hosties and Garbos: A look behind diminutives and pejoratives in Australian English. *Language and civilization: A concerted profusion of essays and studies in honour of Otto Hietsch*, 166-184.

Mester, R. A. (1990). Patterns of truncation. *Linguistic Inquiry, 21*(3), 478-485.

Nathan, G. S. (2008). *Phonology : a cognitive grammar introduction*. Amsterdam ; Philadelphia: John Benjamins Pub. Co.

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. *Vocabulary: Description, acquisition and pedagogy, 14*, 6-19.

Newman, P., & Ahmad, M. (1992). Hypocoristic names in Hausa. *Anthropological Linguistics*, 159-172.

Newmeyer, F. (1988). Extensions and implications of linguistic theory: an overview. *Linguistics: the Cambridge survey, 2*, 1-14.

Newmeyer, F. J. (1983). *Grammatical theory: Its limits and its possibilities*: University of Chicago Press.

O'Grady, W. (2008). The emergentist program. *Lingua, 118*(4), 447-464.

Owens, J., & Ratcliffe, R. R. Morphology. In.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition and contrast. Frequency and the emergence of linguistic structure, ed. by Joan Bybee and Paul Hopper, 137-57. In: Amsterdam: John Benjamins.

Pinker, S. (1999). *Words and rules : the ingredients of language*. London: Weidenfeld & Nicolson.

Pinker, S. (2008). *The stuff of thought : language as a window into human nature*. London: Penguin.

Plag, I. (2003). *Word-formation in English*. Cambridge: Cambridge University Press.

Plénat, M., & Roché, M. (2001). *Prosodic constraints on suffixation in French.* Paper presented at the Topics in Morphology. Selected Papers from the Third Mediterranean Morphology Meeting,(Barcelona.

Prunet, & Idrissi. (2014). Overlapping morphologies in Arabic hypocoristics. *The Form of Structure, the Structure of Form: Essays in honor of Jean Lowenstamm, 12*, 177.

Prunet, J. F. (2014). *Overlapping affixes in Arabic hypocoristics*. Paper presented at the 25th International Congress of Onomastic Sciences.

Rácz, P., Pierrehumbert, J. B., Hay, J. B., & Papp, V. (2015). Morphological emergence. *The Handbook of Language Emergence*, 123-146.

Rainer, F. (2005). Constraints on Productivity. In P. Štekauer & R. Lieber (Eds.), *Handbook of word-formation* (pp. 335-352). Dordrecht: Springer Netherlands.

Ratcliffe, R. R. (2004). Sonority-Based Parsing at the Margins of Arabic Morphology: In Response to Prunet, Beland, and Idrissi (2000) and Davis and Zawaydeh (1999, 2001). *al-'Arabiyya*, 53-75.

Robertson, D. D. (2004). On the Form of Chamorro Hypocoristics. *Working Papers of the Linguistics Circle, 18*(1), 84-94.

Robins, R. H. (1997). *A short history of linguistics* (4th ed. ed.). London: Longman.

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current directions in psychological science, 12*(4), 110-114.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 1926-1928.

Salaberri Zariategi, P. (2003). On hypocoristic formation in Basque. *Fontes Linguae Vasconum, 93 (2003), 329-336.*

Scalise, S. (1986). *Generative morphology* (Vol. 18): Walter de Gruyter.

Scalise, S., & Guevara, E. (2005). The lexicalist approach to word-formation and the notion of the lexicon. *Handbook of word-formation*, 147-187.

Schane, S. A. (1973). *Generative phonology*. Englewood Cliffs ; London: Prentice-Hall.

Scherer, C. (2015). 103. Change in productivity. In *Word-FormationAn International Handbook of the Languages of Europe*.

Sellers, P. (2016). Space, Climate Change, and the Real Meaning of Theory. Retrieved from https://www.newyorker.com/tech/elements/space-climate-change-and-the-real-meaning-of-theory

Shimron, J. (2003). Semitic languages: Are they really root-based? *LANGUAGE ACQUISITION AND LANGUAGE DISORDERS, 28*, 1-28.

Skousen, R. (1989). *Analogical modeling of language*. Dordrecht ; London: Kluwer Academic.

Skousen, R. (1992). *Analogy and structure*. Dordrecht ; London: Kluwer Academic Pub.

Skousen, R., Lonsdale, D., & Parkinson, D. B. (2002a). *Analogical modeling : an exemplar-based approach to language*. Amsterdam: John Benjamins.

Skousen, R., Lonsdale, D., & Parkinson, D. B. (2002b). *Analogical modeling: An exemplar-based approach to language* (Vol. 10): John Benjamins Publishing.

Smith, A. D. M. (2014). Models of language evolution and change. *Wiley Interdisciplinary Reviews: Cognitive Science, 5*(3), 281-293. doi:10.1002/wcs.1285

Spencer, A. (1991). *Morphological theory : an introduction to word structure in generative grammar*. Oxford: Blackwell.

Starks, D., & Taylor-Leech, K. (2011). Research project on nicknames and adolescent identities. *New Zealand Studies in Applied Linguistics, 17*(2), 87.

Stonham, J. T. (1994). *Combinatorial morphology*. Amsterdam: John Benjamins.

Strandquist, R. (2004). Mauritian Creole Hypocoristic Formation. *Working Papers of the Linguistics Circle, 18*(1), 95-106.

Su, D. (2016). Grammar emerges through reuse and modification of prior utterances. *Discourse Studies, 18*(3), 330-353.

Summerell, O. F. (1995). Philosophy of proper names. *Eichler, Ernst & Hilty, Gerold &*.

Taylor, J. R., & Kennedy, R. (2015). Nicknames. In The Oxford Handbook of the Word. Oxford: Oxford University Press. 650-668.

Tomasello, M. (2005). *Constructing a language : a usage-based theory of language acquisition*. Cambridge, Mass. ; London: Harvard University Press.

Ussishkin, A. (2011). Tier segregation. *M. van Oostendorp, C. Ewan, E. Hume, & K. Rice, The Blackwell Companion to Phonology*, 2516-2537.

Van Langendonck, W. (2007). *Theory and typology of proper names* (Vol. 168): Walter de Gruyter.

Vom Bruck, G., & Bodenhorn, B. (2006). *An anthropology of names and naming*: Cambridge University Press.

Wang, Q. (2004). A Prosodic Analysis of Dysyllabicity in Chinese Hypocoristics. *Working Papers of the Linguistics Circle, 18*(1), 107-121.

Watson, J. C. (2006). Arabic morphology: diminutive verbs and diminutive nouns in San’ani Arabic. *Morphology, 16*(2), 189-204.

Watson, J. C. E. (2002). *The phonology and morphology of Arabic*. Oxford: Oxford University Press.

Wierzbicka, A. (1986). Does language reflect culture? Evidence from Australian English. *Language in Society, 15*(3), 349-373.

Witten, I. H. a., Frank, E. a., & Hall, M. A. a. (2011). *Data mining : practical machine learning tools and techniques* (3rd ed. ed.). Burlington, MA: Morgan Kaufmann.

Wong, S.-y. (1997). Nicknames and pet names in Hong Kong. *香港大學學位論文*, 1-0.

Zwicky, A. M., & Pullum, G. K. (1987). Plain Morphology and Expressive Morphology. *1987*, 11. doi:10.3765/bls.v13i0.1817