Open Linguistics 2015; 1: 458–466 DE GRUYTER OPEN

Research Article **Open Access**

Glenn Hadikin

Lexical selection and the evolution of language units

DOI 10.1515/opli-2015-0013
Received September 15, 2014; accepted May 27, 2015

Abstract: In this paper I discuss similarities and differences between a potential new model of language development - lexical selection, and its biological equivalent - natural selection. Based on Dawkins' (1976) concept of the meme I discuss two units of language and explore their potential to be seen as linguistic replicators. The central discussion revolves around two key parts - the units that could potentially play the role of replicators in a lexical selection system and a visual representation of the model proposed. draw on work by Hoey (2005), Wray (2008) and Sinclair (1996, 1998) for the theoretical basis; Croft (2000) is highlighted as a similar framework. Finally brief examples are taken from the free online corpora provided by the corpus analysis tool Sketch Engine (Kilgarriff, Rychly, Smrz and Tugwell 2004) to ground the discussion in real world communicative situations. The examples highlight the point that different situational contexts will allow for different units to flourish based on the local social and linguistic environment. The paper also shows how a close look at the specific context and strings available to a language user at any given moment has potential to illuminate different aspects of language when compared with a more abstract approach.

Keywords: memes, natural selection, lexical priming, needs only analysis, NOA, lexical unit, lexical selection, linguemes

1 Introduction

In 1976 Richard Dawkins changed the way many look at biological evolution with his popular account of a gene's eye view of natural selection – The Selfish Gene. He begins by describing how certain molecules in the early earth's biosphere must have formed *replicators*: rare molecules that can replicate themselves by acting as a mould and attracting smaller molecules in such a way that the larger molecule essentially creates a copy of itself. (Dawkins reminds us that the molecules are not, of course, conscious of this but terms like 'creating a copy of itself' are simply useful shorthand.) DNA is the modern equivalent. This paper discusses the idea that language chunks can be seen as linguistic replicators. It begins with a short review of the literature before moving on to the question of what units might play such a role. A visual representation of my lexical selection model[1] is introduced in section two before I move on to a discussion of specific language strings in a corpus of computer linguistics papers. New terminology is proposed as part of the model. The rest of the introduction will introduce key ideas from both biological and linguistic evolution.

After introducing the concept of physical replicators in biology, Dawkins introduces *memes* as their cultural equivalent:

1 Note that the focus of the paper is *lexical selection* and, as such, will not extend to a detailed discussion of phonology though it is acknowledged this is an important aspect of selection that has been discussed elsewhere (Baxter *et al*. 2009 is an excellent example).

***Corresponding author: Glenn Hadikin:** University of Portsmouth, Portsmouth, United Kingdom. E-mail: glenn.hadikin@port.ac.uk

> The new soup is the soup of human culture. We need a name for the new replicator, a noun which conveys the unit of cultural transmission, or a unit of *imitation*. 'Mimeme' comes from a suitable Greek root, but I want a monosyllable that sounds a bit like 'gene'. I hope my classicist friends will forgive me if I abbreviate mimeme to *meme*. If it is any consolation, it could alternatively be thought of as being related to 'memory' or to the French word *même*. It should be pronounced to rhyme with 'cream'.
>
> Dawkins (1976:192)

Memes then are not physical entities but, as originally formulated, abstract concepts that exist in the minds of people (and arguably a number of other species such as songbirds). They replicate by being shared in a community and mutate by way of small changes. Dawkins gives the examples of tunes, catchphrases and ways of making pots among others.

It is the linguistic side of Dawkins' argument that sets up my own thesis for this paper. I argue that strings of language items – let us imagine words for the moment – enter into a similar competitive environment with other strings and that minor variations will allow an evolution-like mechanism to operate. Whether this linguistic mechanism closely follows modern models of evolutionary theory remains to be seen; the discussion is only just beginning. Note that my argument is not fundamentally different from Dawkins' concept of memes but that it is argued specifically for human language strings without carrying along any assumptions about songbirds or pot-making. Dennett (2003) reminds us that the science of *memetics* is still controversial – seen by critics as just a tool or metaphor that cannot be made literal when applied to all of human culture so this paper should be seen as strictly a linguistic argument. It is important to clarify at this point one key argument I am not making. I am not arguing that the movement and selection of linguistic replicators is necessarily very close to its biological equivalent. In biology, for example, researchers are beginning to understand that the gene does not play as central a role in evolutionary theory as was once assumed (see arguments in Jablonka, Zeligowski and Lamb 2014 for example). Developments in evolutionary theory should, however, not be seen as a driving force for linguistic theory. Lexical Selection may one day turn out to be notably different in its detail.

This is certainly not the first paper to discuss the similarities between language replicators and biological replicators. Pagel (2009) provides an excellent overview of many of the similarities from a biologist's point of view and he reminds us that many of them were, in fact, suggested by Darwin himself (Darwin 1871). Pagel suggests that words, phonemes and syntax[2] are examples of discrete units and that they replicate by means of teaching, learning and imitation. He also offers ancient texts as the equivalent of fossils and language death in relation to extinction. The work focuses largely on the word as the unit of analysis by using Swadesh's (1952) list of two hundred common words such as body parts, pronouns and numerals that could be expected in many languages. The work shows that researchers can produce matrices with the number of 'meanings' on one side against the number of languages that have that meaning in a selected set. This allows for statistical analyses and comparisons with gene sequence matrices. Pagel's discussion includes a surprising association between the distribution of animal species and the distribution of human languages when plotted against latitude.

Kirkby and Hurford's (2002) Iterated Learning Model (ILM) moves closer to an actual model of language change; it is described as "an adaptive system that operates on a timescale between individual learning and biological evolution" (Kirkby 2007:1) and is well-supported by computer and mathematical models. Such models are, however, necessarily abstracted away from real speakers and actual utterances with work based on idealised languages. This is important work that highlights the potential of an evolutionary approach but its distance from actual utterances can seem divorced from work in text analysis and corpus linguistics which focus more on the behaviour of strings of lemmas and wordforms. As we move to a complete theory of language we will need work at both ends of the abstract-utterance continuum. An evolution-like theory that explicitly seeks to work at the level of utterances is Croft's (2000) Theory of Utterance Selection (TUS).

Croft's work is similar to the way I conceptualise lexical selection. It is inspired by Hull's (1988) work on a generalised theory of selection for all evolution-like processes and, as such, considers the nature of potential replicators in some detail. Croft suggests "embodied linguistic structures, anything from a phoneme

2 Pagel lists 'syntax' as a possible unit; one can assume he means syntactic structures.

to a morpheme to a word to a syntactic construction, and also their conventional semantic/discourse-functional (information-structural) values" (2000:28); he calls these units *linguemes*. Linguemes then are not necessarily individual words but many examples Croft uses either imply that they tend to be or he writes at a level somewhat removed from texts so that different readers may walk away with different views of what a lingueme actually looks like when or, indeed, if one can exist on a page. In section two I will discuss the extent to which an *extended selection unit* (ESU) in my terminology should be seen as equivalent to a lingueme.

2 What units?

For this paper I will focus on two units that could potentially play a key role in the replication system – *words* and *extended lexical units*. Words are arguably the most logical starting point for a discussion of language units. As Hoey (2005) reminds us, the dictionaries and thesauruses of the 18th and 19th centuries reflect the classical idea of the word with information about pronunciation, grammar and etymology. The concept of *word* is, however, rather woollier than it may first appear. The simple distinction between *types* and *tokens* in corpus studies highlights the first point of concern. If one considers the string *me me me* we would say it consists of three tokens (occurrences of) a single type – the form *me*. Lemmatisation further complicates matters. Hoey cites Williams (1998), for example, who shows the collocational behaviour of *gene* is quite distinct from that of *genes*. If we are to consider strings of words moving through a community it would make sense to read this as strings of wordforms in the first instance; in many cases, however, a concept I call the language user's *lemma space* comes into play. I will use the string *pen drive* as an example that many of us may remember using for the first time. If, like me, an adult came across the string for the first time she might receive the form *pen drive* as input but then may immediately map that form onto its lemma space and be in a position to produce other wordforms such as *pen drives* at the first opportunity for producing output. Thus both the wordform and its lemma space could potentially play an important role as the unit moves through a community. The hedging in my last two sentences is because of Wray's (2002) concept - Needs Only Analysis (NOA) - which I discuss in a moment.

Despite the complexities of the concept WORD it is seen by millions, if not billions, of everyday language users as an important unit and this will have a powerful influence on the development of languages that we are just beginning to understand. The more people beyond academia become familiar with concepts such as collocation and multi-word units, the more likely students will study and become familiar with units longer than the word thus affecting selection processes in ways we are yet to understand; this 'public perception' aspect of language selection has been understated in previous work.

Wray perceives the word as the "fundamental currency of processing" (Wray 2008:67) in Hoey's model and states that this is a substantial difference between Hoey's *Lexical Priming* theory and her Needs Only Analysis (NOA) model; in NOA Wray convincingly argues that in a first language acquisition context children do not break down language strings into their smallest components but, in fact, store the whole string unless there is a specific need for further analysis. To further explore this apparent conflict I need to say a little about Hoey's concepts of *Lexical Priming* and *nesting*.

Lexical Priming is fundamentally the knowledge a language user has about a language unit such as a word or phrase – it includes collocation, information about grammatical patterns (colligation) and semantic and pragmatic patterns as well as being genre and domain specific. One of Hoey's examples is the sentence

> In Winter Hammerfest is a thirty-hour ride by bus from Oslo, though why anyone would want to go there in winter is a question worth considering.
>
> Hoey (2005:5)

Hoey gives this as an example of natural primings for many English-speaking readers – it is taken from a travel book by Bill Bryson. As an exercise I have put this sentence up on a screen for students to think about for a few seconds and then blacked out the screen before asking students to reproduce the sentence again as accurately as possible. It is noteworthy that many British students who have tried this activity produced

the string *bus ride* rather than *ride by bus*. Indeed the British National Corpus (BNC)[3] has 47 occurrences of *bus ride* compared with none for *ride by bus* suggesting that British speakers are more strongly primed for the former string. Though it is ultimately the human speaker that is primed for a certain word use, Hoey talks of words being primed as a kind of shorthand. He might say, for example, that the word *bus* appears to be primed for collocation with *ride* in this example but it is always the human language users that are primed for use in that context.

Nesting is a situation where a combination such as *bus ride* takes on primings which may differ from its constituent words *bus* and *ride*. Hoey clearly states "I have focused on the word as a convenient starting point for the description of priming, rather than for theoretical reasons" (Hoey 2005: 158) so I would argue that his model is not notably at odds with Wray's view. Similar to *nesting* Wray argues for a *heteromorphic lexicon* in which morphemes, words and multi-word expressions exist separately in the mental lexicon but can reuse constituent components; *bus* and *bus ride* would make up two entries rather than require a language user to construct the longer string every time (Wray 2008). The mental lexicon would, however, play a less central role in a model of lexical selection.

A lexical selection model would involve a complex of humans and texts in a society. Humans begin to learn a string such as *bus* with basic primings – at any stage of the acquisition process the user may prepare to reproduce the string in a social context (see figure 1). This gives two potential results; either the string has been stored and understood by the user in a form that closely matches the original input – say, for example, a pronunciation form that is accepted by the community and is understood with an appropriate set of primings that would allow it to be used without correction. Or, alternatively, it is stored in an altered form. It could sound notably different or be stored with a set of primings that lack information about its typical colligational relationship with articles, for example. Note that *primings* is used throughout this paper in the sense described in Hoey (2005) and, thus includes conscious memorisation processes as well as subconscious effects. Social pressure would act on the whole system and influence the language user at every stage. The second level of figure one that shows the two forms of lexical storage closely reflects Wray's (2002, 2008) work so 'string memorised in identical form' would include both formulaic word strings and strings of non-formulaic language in cases where the individual is strongly primed to produce the standard output for a particular situation in that community. 'string memorised in altered form' would consist of strings where the individual is more weakly primed to produce the standard output for any given situation.

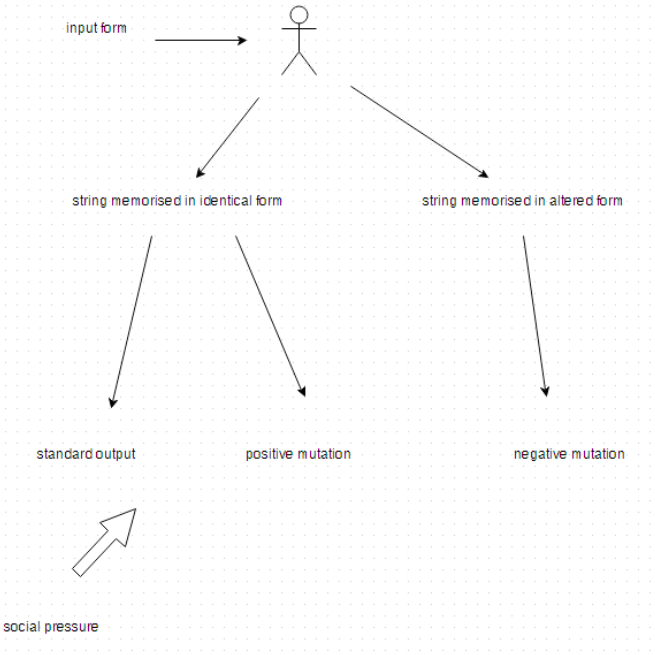

Figure 1: Lexical selection process for any given string

3 BNC data was accessed via http://www.sketchengine.co.uk/

The user with a (mentally) stored form that is highly congruent with the original can choose to reproduce the string in that form or to be creative and consciously adjust the primings of the next listener - a *positive mutation*. If the string is stored in an altered form the speaker will be seen to have less control over the language by the local community and would more likely produce a *negative mutation* – this would likely be seen as a mistake by the community and would be at high risk of correction or deletion (and hence possible removal from a lexical gene pool unless the speaker adopts the form nonetheless). Wray (personal communication) points out that mutations in such a lexical selection model are not random in the same way as biological mutations are often seen to be; this view of biological mutation is not necessarily an accurate one, however (see Martincorena, Seshasayee and Luscombe 2012, for example). It is important to note that important differences do exist between biological evolution and language and this must be taken into account when exploring the model further.

Words do not occupy a privileged place in this description. Like their role in Lexical Priming (Hoey 2005) and Wray's heteromorphic lexicon (Wray 2002) they are just short strings that often make convenient starting points for analysis. (The public view of the word as a unit is, however, a very important contextual factor). In corpus linguistics we often begin by searching for a node such as *bus* and then discussing the linguistic environment around it by means of concordance lines (figure two shows an example).

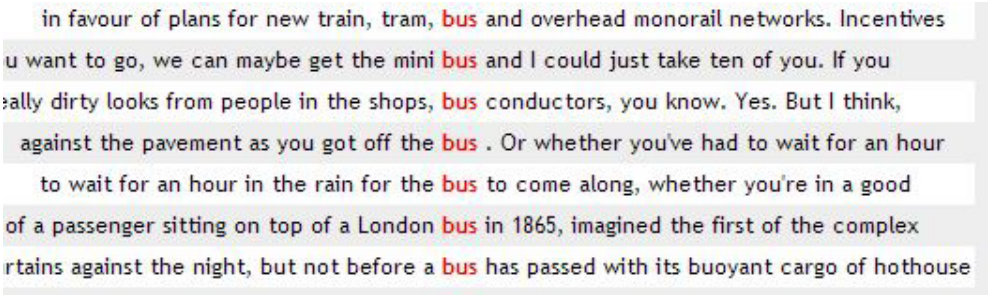

Figure 2: Sample concordance of node *bus* from BNC/Sketchengine

It is interesting to note that a literal sense of the word *node* is a knot in cloth ('node' 2014) which provides a useful visual metaphor. The node is not the whole unit but provides a handle for us to grasp and begin to look at the rest of the material. As long as language learners and teachers perceive words as important units words will be acted upon and treated as selection units but in naturalistic settings and first language contexts we must consider a wider range of *lexical units*.

Sinclair's (1996, 1998) model of the *extended lexical unit* tells us that each unit consists of a *core* as well as *collocational* and *colligational patterns*, *semantic prosody* and *semantic preference*. Semantic prosody relates to the language user's reason for choosing a given expression so, in Sinclair's example, *naked eye* has a semantic prosody of *difficulty; semantic preference* relates to the semantic area of the surrounding lexis. In a sense, Hoey's Lexical Priming (2005) argues for the same units but adds a textual dimension in that any unit (words or longer strings) might be primed to form cohesive patterns. Hoey also conflates the concepts of semantic prosody and semantic preference replacing them with *semantic association*. He defines semantic association as "when a word or word sequence is associated in the mind of a language user with a semantic set or class" Hoey (2005:24).

It is likely that a form of extended lexical unit – let us call them *extended selection units* or ESUs – would play a role in a lexical selection process and the smaller units such as words and morphemes would vary within them as users speak, write and help the ESU move through a community. I will call the smaller units *variators*. From the top-down perspective of memes and replicators an ESU is equivalent to a lingueme (if one assumes strings not exemplified by Croft such as n-grams and lexical bundles are subject to the same mechanisms) and lexical selection would be a mechanism within the general framework of Croft's Theory of Utterance Selection (Croft 2000); one must always remember, however, that this does not pin linguistic theory to biological evolutionary theory. The two theories must be allowed freedom to develop

independently. A potentially important difference is at the bottom-up level of actual utterances. Croft's theory either assumes a traditional view of phrase structure and syntax or leaves the interpretation of what can and what cannot be a lingueme open to the reader. An ESU is more closely aligned to Hoey's Lexical Priming (Hoey 2005) in the sense it could potentially be any string of language. This leaves room for potential units such as those indicated by the written forms *of the* or *in the* to be analysed and discussed as ESUs depending on the primings and behaviour of a speech community; their status as a lingueme in Croft's work is less clear.

3 Movement and competition

To explore the idea of how an ESU might move through a community I will begin with the string *the accuracy of the* in the open corpora provided by Sketch Engine (see Kilgarriff, Rychly, Smrz et al. 2004 for details). This string was not selected as a particularly unusual piece of language - in many ways it is completely unexceptional. I chose the string as I was exploring the issue of whether a frame - *the * of the* - could meaningfully be seen as a selection unit; I also felt it was important to avoid a well known formulaic expression as these could be seen as extreme cases. As no language user ever needs a frame without a communicative situation I chose, somewhat arbitrarily, a 4-word string that was well-represented in the data to allow me to move towards more concrete examples of language use.

It occurs in all four of the free corpora available via Sketch Engine - the ACL is an archive of computer linguistics publications, British Academic Spoken English Corpus (BASE), British Academic Written English Corpus (BAWE) and Brown, a corpus of general US English prepared in 1964, as shown in figure 3.

ACL	1130
BASE	1
BAWE	91
Brown	5

Figure 3: Raw frequency of *the accuracy of the* in four corpora

Let us consider example (1) from the ACL corpus:

(1) We show in detail our findings about syntactic levels (how often graph matching helped assign a relation between two clauses, a verb and its arguments, or a noun and its modifier) and about the accuracy of the suggestion.

The writer is likely to be clear that they want to refer to the concept of accuracy at this point and will be primed to use the * *of the* * frame in formal writing. Indeed this writer produces *the requirements of the domain*, *the behaviour of the system*, *the analysis of the input text*, *the end of the experiment* and *the structures of the syntactic units* in the same paper.

the	accuracy	of	the	suggestion
1/1	1/2	1/1	1/1	1/1130
100	50	100	100	0.09

Figure 4: Chart showing the percentage that each lexical item fills the frame in ACL (based on Biber 2009)

The researcher is also likely to be primed to use the string *the accuracy of the* * in such a situation. Figure four shows the number of times each slot in this structure is occupied by the given word; the word *accuracy*, for example, occurs once in two occurrences of *the * of the suggestion*. A second usage in the corpus is *the quality of the suggestions* in a different text. This shows us the relatively fixed nature of *the accuracy of the* *

as a frame with the last slot remaining very open for the writer or speaker to insert the appropriate selection for the context (in a practical example, of course, the writer would be rather more limited depending on what system or concept is being described).

In a sample of 300 lines the most frequent complement for this frame was *parser* so I used the Corpus Query Language (CQL) term "accuracy" []{0,2} "parser" on the full corpus to explore any cases when a writer may be faced with a genuine choice of terms when they need to describe the accuracy of a specific parser.

of the parser was 92.8%, higher than the	accuracy of a parser	based on an SLM without ACTs (89.8%). This
ictionary built with the system improves the	accuracy of a parser	by an appreciable amount'. 1 Motivation
nate that function. We can measure the (in)	accuracy of the parser	by the amount of additional information
a parse is finished, the efficiency and	accuracy of the parser	can be radically altered. The parser orders
in order to improve the general parsing	accuracy. The parser	computes true Viterbi parses unlike most
when it is heavily biased toward nouns, the	accuracy of the parser	depends on having some verbs. After pre-processing
have a broad- coverage parser that has the	accuracy of a parser	designed specifically for a domain. One
which usually require high accuracy. The	accuracy of the parser	directly affects the accuracy of the generated
is clear that, with respect to unlabeled	accuracy, our parser	does not quite reach state-of-the-art performance
ion of Penn Treebank corpus. Currently the	accuracy of the parser	drops down to 82% on GTB-beta, and although
xception. The first row of table 4 shows the	accuracy of the parser	for different arc lengths under the baseline
for real-world applications. However, the	accuracy of the parser	in terms of dependency analysis was significantly
Category Prediction Table 2 reports the	accuracy of the parser	in the empty category (EC) prediction task
given parse, and increases the parser's	accuracy. The parser	introduces syntactic constraints that perform
From the 1-best result we see that the base	accuracy of the parser	is 89.7%.1 2-best and 10-best show dramatic
d dependencies shows that the dependency	accuracy of the parser	is 90.1% on child language transcripts
insertions (Ins), the overall labelling	accuracy by the parser	is around 80%. 4.3 Bracketing Accuracy
from a reference standpoint? Currently, the	accuracy of the parser	is couched in syntactic terms. The precision
text. Given the difficulty of the task, the	accuracy of the parser	is encouraging. The errors made by the
corpus being non-projective .2 Since the	accuracy of our parser	is far from 98.2 %, it is useful to require
searched using the parsing model. Thus the	accuracy of the parser	is greatly reduced, as shown in rows 2
result. However, generally speaking, the	accuracy of Japanese parser	is low compared with that of Japanese morphological
to label edges with semantic types. The	accuracy of their parser	is lower than that of Yamada and Matsumoto
both on the simple parsing task, where the	accuracy of the parser	is measured on the standard Parseval measures
both on the simple parsing task, where the	accuracy of the parser	is measured on the standard Parseval measures
ord and 2-word sentences are very few. For	accuracy, the parser	is more effective for shorter sentences
98.6% of the sentences in section 23. The	accuracy of the parser	is reported in Clark and Curran (2004b)
suffices like -1y, -y, -ed, -d, and -s. The	accuracy of this parser	is reported Proceedings of EACL '99 Synonyms
example, adding a new syntactic rule. The	accuracy of the parser	is significantly improved, from 69.4% to
, 0.01, and 0.5, which represent trigram	accuracy, parser	model accuracy, and mixed accuracy. Surprisingly
current choice of A. We then determine the	accuracy of the parser	on a held-out development set using the
h(i) for all i in this leaf termine the	accuracy of the parser	on a held-out development set using the

Figure 5: Sample concordance showing results of CQL results for "accuracy" []{0,2} "parser" in the ACL corpus/Sketch Engine

Figure five shows a sample of the results. It is clear *the accuracy of a parser* would not be a suitable string for a writer who wants to refer to a specific case - it has a more general meaning. The string *the accuracy of the parser*, it seems, would most likely compete with *the accuracy of our parser* and *the accuracy of their parser* depending on the ownership of the parser. Let us assume a situation where the writer was discussing a parser they had developed. This would leave *the accuracy of the parser, the accuracy of this parser* and *the accuracy of our parser* as remaining options. Priming effects will influence the likelihood of any one form being chosen at each reference point - this includes the writer's concern to avoid repetition in their writing.

A careful study of the writer's papers and spoken presentations would give us a fascinating insight into his or her lexical primings and the way the writer treats each string as a unit - this would make an interesting PhD project but to fully explore lexical selection we must turn to transmission and encourage future studies where at least a second speaker or writer carries the linguistic information to a new audience.

The strings *the accuracy of the* and *the accuracy of the parser* were partially chosen as challenging cases that would force us to consider weaknesses in this view of language. There is little doubt that formulaic expressions such as idioms and restricted collocations are being shared by communities (partially) because they mark a speaker as being a competent member of the language community in question but the strings chosen here are less salient as chunks and may be broken up by language learners. Recall the visual representation of the lexical selection model in figure one. By writing this paper I have become primed to use the string *the accuracy of the parser* as a fixed string; I may have primed some readers to also recall it as a formulaic string. Our next recorded usage of the string or, indeed, our attempt at using a similar string to express the concept would provide a useful indicator of how it has moved as a unit. Most formal writing would comprise examples of standard output in that the string would be produced in an identical form. Clearly an altered form such as *the accuracy of this parser* would be a reasonable choice in writing - in this context this would comprise a positive mutation. A negative mutation in such a scenario might result in the string *the accuracy of parser* and would be at high risk of intervention from a teacher or editor before going out as a second generation string. The string *the accuracy of Japanese parser* from Figure 5 is a good example of how a string must be suited to its environment to spread. In a British or American journal, for example, one might imagine an editor suggesting a change and preventing such a string from spreading through the community but in a New English environment such as Japanese or Korean English this string may be accepted and move freely around its community (Hadikin 2014 provides several examples of this).

4 Conclusion

In this paper I introduce a way of looking at language units that will be new to some fields of language study. Based on Dawkins' concept of memes (Dawkins 1976) I have looked at the movement and structure of two potential language replicators - the word and Sinclair's suggestion of an extended lexical unit (Sinclar 1996, 1998). Hoey's (2005) *Theory of Lexical Priming* and Wray's (2008) concept of the *heteromorphic lexicon* remind us that words do not appear to move as independent entities and, building on key arguments from each model, I propose a visual representation of a potential lexical selection process with two kinds of 'mutation' - one based on a language user storing the language string in a form highly congruent with input form and then altering the structure in a way that is likely to spread successfully - and a second that would likely be seen as erroneous in by the local language community.

Compared with single words, Sinclair's units make a better candidate for the kind of unit that would act as replicators because they allow for the kind of data patterns described in Hoey (2005) and Wray's (2008) arguments and, for ease of analysis and further discussion, I propose new terminology to reflect the concept of genes and bases - extended selection units (ESUs) and variators respectively. Finally, I briefly discuss the movement and selection involved when researchers and students use the strings *the accuracy of the* and *the accuracy of the parser* in free online corpora. The examples remind us that we cannot work entirely with abstractions in a paper such as this one because to truly visualise language users reading an ESU and attempting to reproduce it in a second generation context we must consider a real world example and a real world language community. The need for future empirical work is clear - as well as corpus studies and the kind of formulaic language work described in Wray (2008) I propose studies that track how particular strings move between one user and the next. Such studies will give us a firmer understanding of what sort of linguistic information actually gets carried as an ESU so that one can more confidently explore the extent of similarity between Sinclair's abstract extended lexical units and the ESUs that actually move from text to human and from human to text. There is also a need for new computer models that help researchers explore the information carried by a number of different units rather than rely on single words as the unit of analysis.

References

Baxter, Gareth J., Richard A. Blythe, William Croft et al. 2009. Modeling language change: an evaluation of Trudgill's theory of the emergence of New Zealand English. *Language Variation and Change* 21(2). 157-196.

Biber, Douglas. 2009. A corpus-driven approach to formulaic language in English: multi- word patterns in speech and writing. Paper presented at Corpus Linguistics 2009, University of Liverpool, 20-23 July.

Croft, William. 2000. *Explaining Language Change*. Harlow: Pearson Education.

Darwin, Charles. 1871. *The descent of man*. London: Murray.

Dawkins, Richard. 1976. *The Selfish Gene*. Oxford: Oxford University press.

Dennett, Daniel. 2003. *Freedom evolves*. New York City: Viking.

Hadikin, Glenn. 2014. *Korean English: a corpus driven study of a New English*. Amsterdam & Philadelphia: John Benjamins.

Hoey, Michael. 2005. *Lexical Priming: a new theory of words and language*. London: Routledge.

Hull, David. 1988. *Science as a process: an evolutionary account of the social and conceptual development of science*. Chicago: University of Chicago press.

Jablonka, Eva, Anna Zeligowski & Marion J. Lamb. 2014. *Evolution in Four Dimensions : Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*. Cambridge, MA: MIT press.

Kilgariff, Adam, Pavel Rychly, Pavel Smrz et al. 2004. The Sketch Engine. In Proceedings of Euralex 2004, Lorient, France, 6- 10 July, 105-116.

Kirby, Simon. 2007. The Evolution of Meaning-space Structure through Iterated Learning. In Proceedings of the Second International Symposium on the Emergence and Evolution of Linguistic Communication, University of Hertfordshire, 12-15 April, http://www.lel.ed.ac.uk/~simon/Papers/Kirby/kirby%20aisb%20v2.pdf (accessed 5th April 2015).

Martincorena, Iñigo, Aswin Seshasayee & Nicholas Luscombe. 2012. Evidence of non- random mutation rates suggests an evolutionary risk management strategy. *Nature* 485 (7396). 95-98.

node. In *Oxford English Dictionary*. 2014.

Pagel, Mark. 2009. Human language as a culturally transmitted replicator. *Nature reviews*: *genetics* 10. 405-415.

Sinclair, John. 1996. The search for units of meaning. *Textus* 9. 75-106.

Sinclair, John. 1998. The lexical item. In Edda Weigand (ed.) *Contrastive Lexical Semantics*. 1-24. Amsterdam & Philadelphia: John Benjamins.

Swadesh, Morris. 1952. Lexico-statistic dating of prehistoric ethnic contacts. In Proceedings of the American Philosophical Society 96(4). 453-463.

Williams, Geoffrey. 1998. Collocational networks: interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics* 3(1). 151-171.

Wray, Alison. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Wray, Alison. 2008. *Formulaic Language: pushing the boundaries*. Oxford: Oxford University Press.