Vsevolod Kapatsinski

Creativity through Inhibition (of the First Production that Comes to Mind)

Abstract: This paper proposes that, from a speaker-internal mechanistic perspective, creativity results from the suppression of a prepotent production activated by the message and the context in which the speaker is trying to express it. It illustrates a recent proposal for how such suppression might be accomplished by the production system, the Negative Feedback Cycle (Kapatsinski 2022) by a few examples. The Negative Feedback Cycle suppresses the production of forms that are likely to have unintended consequences before they are blurted out. The Negative Feedback Cycle's main function is to avoid overextension of frequent forms. However, as a side effect, it generates avoidance behaviors, such as avoiding the production of forms that are likely to be misunderstood. And, as another, more pleasant side effect, it also generates many creative linguistic behaviors that seed linguistic change, including backformation and circumlocution.

1 Introduction

A contestant stands on the stage of a popular American TV show. He is sweating. Ten thousand dollars are on the line. The question, part of the third-grade curriculum: "What is the singular form of the word *lice*?" The man stammers. The host comments "I didn't even know there is a singular form. I thought they travel in packs." After several minutes of embarrassment, the contestant finally volunteers, uncertainly, the guess *lie* (perhaps, on analogy with *die-dice*). Also guessing are four professional cheerleaders. They are not competing for money. They converge immediately on a different answer, also prescriptively wrong but far more obvious – *lice*. (Video available at https://youtu.be/sGKuNhQ7uRA?t=968).[1]

The present paper asks the following question: given the immediate availability of the form *lice* provided by the host, how does the contestant decide not to

1 As shown by Bybee and Slobin (1982) such overextensions are the most persistent type of error in morphological production (see also Harmon et al., 2023; Hoeffner and McClelland, 1993; Taatgen and Anderson, 2002). They are also common in language change where they are likewise predictable from frequency and semantic similarity (Brochhagen et al., 2023; Bybee and Brewer, 1980; Tiersma, 1982).

https://doi.org/10.1515/9783111321905-004

use that form to express the related singular meaning (like all four cheerleaders did)? In other words, how does the contestant avoid producing *lice* immediately – since it is clearly the most accessible form in the moment, having been provided to him by the host – and instead continue painstakingly searching for a better answer? That is, how does he avoid overextending the form *lice* to the singular meaning LOUSE? The painstaking search for a less obvious expression of the intended meaning is, I argue, a basic prerequisite for creativity.

The intuition I pursue in this paper is that *lice* is not produced because it is a good cue to a meaning that the speaker does *not* want to express – PLURAL. This intuition is spelled out in a connectionist interactive activation framework for language production (Kapatsinski 2022). I illustrate how this candidate mechanism for suppressing overextension may produce, as a side effect, a number of creative linguistic behaviors.

The proposed framework aims to assume a minimum of linguistic structure. It makes the constructionist assumption that the language system is a network of direct form-meaning mappings (Bybee 2001; Goldberg 1995), where *form* is a surface phonetic representation that mediates between acoustics and meaning in perception and between meaning and articulation in production (Kapatsinski 2021). The proposed framework assumes that there are forms of various sizes, as in interactive activation models (Dell 1985, 1986), though it follows the usage-based tradition in abandoning the assumption that these sizes correspond to traditional linguistic units (Bybee and McClelland 2005; Langacker 1987).

Most importantly, the present framework assumes a connectionist view of the mind, in which the forms and meanings are organized in a network, and processing consists entirely of the spread of activation and inhibition. Under such a view, forms are not either licensed by the language system or lie outside of it – all we have is activation, and activation is a matter of degree.

This has an interesting consequence for the notion of creativity. In an influential recent paper, Sampson (2016), has argued that the simple application of existing constructions (form-meaning mappings) to new input forms (i.e., productivity) is distinct from creativity, or F(ixed)-creativity vs. E(nlarging)-creativity in his terms. (E-)creativity requires *extension* beyond the system, breaking the rules. This distinction presupposes that linguistic generalizations rely on classical categories where an input either is or is not eligible to undergo a particular rule (see also Hoffman 2019).

From a connectionist perspective, forms do not have necessary and sufficient conditions on use. In such a system, forms are activated by distributed semantic / contextual representations. Because these representations are distributed, similar contexts and meanings share cues (activated nodes) – the same nodes participate in representing multiple similar meanings (Hinton et al. 1986). As we shall see,

extension of known forms to new contexts is an inevitable side effect of the distributed nature of these representations for the contexts and cannot be distinguished from following the rules (Bybee and McClelland 2005; see also Suttle and Goldberg 2011, for a related perspective).

Extensions can vary in how similar the original use of a form is to its new use, and in how they are perceived by listeners, but all rely on the same basic mechanism – activation of forms by distributed semantic patterns. From a mechanistic perspective, extensions – no matter how creative-looking – are therefore not true creativity because the producer simply says the first thing that comes to mind in accordance with the normal functioning of the system. Creativity requires following the path less traveled, which I hypothesize requires (conscious or unconscious) reflection on the likely consequences of what one is about to say. The NFC provides a possible implementation for such reflection.

As mentioned earlier, processing consists entirely of the spread of activation and inhibition. Activation and inhibition continue spreading through the network until a contextually-determined deadline, which is influenced by factors like whether one is in danger of losing the floor (Holler et al. 2021) and whether the choice is a consequential one (Nozari and Hepner 2019). Thus, in our original example, longer processing time is allowed by the contestant than by the cheerleaders, because for the contestant thousands of dollars are on the line. This longer processing is, I contend, what allows him to generate a less likely form because it takes time to suppress the most obvious answer and settle on a less obvious one.

In the present framework, the process of suppressing the most obvious answer and settling on a less obvious one is what makes creative production different from routine production – that is, creative productions do go beyond what the language system would normally generate (Sampson 2016), but not because they extend the acceptable range of productions. Rather, it is because they are not the productions that the speaker would most readily generate in the present context. They require time and effort for activation and inhibition to spread beyond the most likely outcome.[2]

Creativity involves producing something new and unexpected, yet appropriate for expressing the intended message (e.g., Heinen and Johnson 2018; Rastelli

2 A demonstration that such networks can produce different outputs depending on the length of time activation is allowed to spread is provided by McClelland (1981), reimplemented by Axel Cleeremans at https://axc.ulb.be/pages/interactive-activation-application. However, the architecture of that network is such that spread of activation will not produce more creative solutions: activation spread in that network suppresses retrieval of unusual facts about the gang members so that they resemble more typical members of the gang with more processing.

et al. 2022). Therefore, in this framework, creativity requires 1) activation to spread through the network from the intended message, so that the intended message affects form selection, and 2) some way to avoid producing the most obvious expression of the intended meaning – the first form that comes to mind – given time to come up with an alternative expression. That is, to do something new in a familiar context, one needs to suppress the familiar actions that the context activates most strongly (Harmon and Kapatsinski 2021). In other words, being creative usually requires overcoming the influence of habit (see also Wood and Neal 2007). I note that creativity in this sense is not necessarily conscious – it might be, but it need not be. That is, there is no causal connection between creativity and consciousness – a system could be creative without being conscious.

Given the above, I adopt the following as the operational definition of creativity: *a creative production is less accessible than some other production(s) given the intended message and the context in which it is expressed, in the moment of production*.[3] The lower accessibility of a creative production means that it is activated by the message and context less strongly and therefore takes longer to come to mind than the more accessible alternative(s) (Oldfield and Wingfield 1965). In this sense, the contestant's effortful production of *lie* as the singular form of *lice* is creative, whereas an immediate production of *lice* as the singular form of *lice* when prompted with the plural form *lice* is not.[4] Though both productions are prescriptively wrong, only one required suppressing a more accessible production.

This definition of creativity is intentionally speaker-internal, as our goal is modeling the functioning of the production system. The production may or may not succeed in transmitting the intended message to the audience, and may or may not look creative to the audience or an outside observer. Though many of the creative productions in the speaker-internal sense do look creative to others,

3 A creative production in this sense is different from an error. First, the erroneousness of a production is in the eye of the beholder. Therefore, errors can be generated in different ways, only some of them creative. Second, many errors are demonstrably not creative because they are quicker than or as quick as correct productions. This means that they involve producing the form that was most accessible in the moment of production. Others are slow, but only because the speaker is in a state of uncertainty about what to produce – multiple alternatives are activated about equally (Staub 2009). Indeed, it appears unlikely that many errors are creative in the present sense, because an error (by definition) matches the intended message less well than an alternative production. So it is unlikely to arise by the speaker detecting that this alternative production doesn't match the message well enough and suppressing it.

4 The contestant puts himself into a tip-of-the-tongue state, which – unlike the usual tip-of-the-tongue state caused by lack of any dominant response – could have been avoided by producing *lice*. The NFC puts him in this state and gets him out of it.

whether they do or don't makes no difference to whether the speaker suppressed their habitual way of expressing their intended message in producing what they produced. This definition is also not intended to subsume all apparently creative linguistic behaviors. For example, sometimes there is no established way to express a message. In this case, processing may not require the proposed mechanism to suppress a more likely expression, but the result can look or sound very creative to an observer.[5] I also do not wish to imply that creativity in this sense is intentionally creative. Indeed, it is often not. The contestant in our original example was not intentionally trying to be creative – he was intending to be as conventional as possible! However, he was (consciously or unconsciously) intending not to blurt out the most accessible form (*lice*). Behaviors that look creative are often an unintended consequence of this intention.

2 Extension: Apparent Creativity

Some behaviors that look creative to an observer arise out of the habitual functioning of the language production system, and therefore are not creative in our sense. Perhaps the best example of this kind is semantic (over)extension (Brochhagen, Boleda, Galdoni and Xu 2023; Gershkoff-Stowe and Smith 1997; Harmon and Kapatsinski 2017; Naigles and Gelman 1995), which (in the present framework) subsumes morphological paradigm leveling (Bybee and Brewer 1980; Harmon et al. 2023; Hoeffner and McClelland 1993; Tiersma 1982).

For example, the child who says *kitty* when presented with a cow is often not truly being creative. Instead, *kitty* is the form activated most strongly by the semantics of a cow: the child either has not yet learned the word *cow* and therefore has no better-matching word in their vocabulary for referring to a CUTE.BOVINE.ANIMAL, or the better-matching word *cow* is simply less frequent than *kitty* and therefore less accessible despite receiving more activation from the intended message – in the child's experience a CUTE.ANIMAL is usually a kitty (Gershkoff-Stowe and Smith 1997; Harmon and Kapatsinski 2017; Naigles and Gelman 1995). Similarly, extending *lice* to mean one louse would not be truly creative because the much greater frequency of *lice* compared to *louse* means that the message LOUSE.SINGULAR is likely to activate *lice* more than *louse* (Harmon and Kapatsinski 2017; Hoeffner and McClelland 1993).

5 This apparent creativity may, perhaps, be based on the listener's simulation of how difficult it would be for them to come up with the same expression for the same meaning. We leave this for future work

Harmon and Kapatsinski (2017) show that such accessibility-driven extensions are not restricted to children and can be elicited experimentally in adults – a form is more likely to be extended to new related meanings if it is frequent in the speaker's prior experience. They further show that there is no preference to extend frequent forms if accessibility differences between frequent and infrequent forms are leveled. This result suggests that extensions result from habit: they are produced because they are more accessible than alternatives, and are therefore accessed before these alternatives come to mind, or (more formally) reach a level of activation needed to be selected for production.

Accessibility-driven extension is illustrated in Figure 1. Here, arrow lengths represent connection strengths, which are proportional to the frequency with which the meaning was expressed by the form in the speaker's experience (see Kapatsinski and Harmon 2017, for a proof that more complex learning algorithms would yield the same result). The dashed line in Figure 1 demonstrates the state of the production system at a point at which the frequent form has been activated by the message (to a level sufficient for production) while the less frequent form has not. An accessibility-driven extension is inevitable at this point, *as long as the speaker starts speaking*.

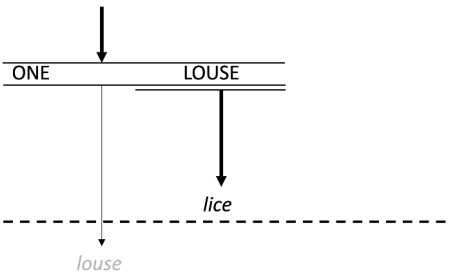

Figure 1: Overextension of *lice* to the meaning ONE LOUSE by a speaker who knows the forms *lice* and *louse* but uses the form *lice* more frequently because lice "usually travel in packs". In this speaker's experience, the form *lice* is much more probable than *louse* given a LOUSE-related message. As a result, *lice* becomes activated more quickly than *louse* (shown by the shorter and thicker arrow) even if the intended meaning activated by the message (rootless top-down arrow at the top of the figure) is ONE LOUSE, and *lice* does not fully match that meaning (matching LOUSE only). The dashed line shows a point in time at which *lice* has already been activated, and *louse* has not yet (which is why it is greyed out). At this point, only *lice* can be produced and overextension therefore appears inevitable.

3 Avoiding Overextension: The Negative Feedback Cycle

To avoid an accessibility-driven extension, production has to be delayed until the less frequent form is activated enough to have a fighting chance against its more frequent competitor. For the speaker to delay production despite having accessed a form, they must sense that there is likely to be something wrong with the form they are about to say, or else they must have an inkling that there is a better option. Otherwise, there is no reason to suppress the most accessible form and delay speaking. Fortunately, speakers can flexibly delay speaking when they have (and sense that they need) time to plan (Holler et al. 2021).

What would make a less frequent form worth waiting for is its superior ability to transmit the intended message to the listener. In fact, speakers who extend frequent forms to new meanings often consider them to be relatively poor expressions of these new meanings. For example, children calling a cow a *kitty* admit that it is not a kitty and would look at a kitty and not a cow when hearing the word *kitty* (Naigles and Gelman 1995). Similarly, Harmon and Kapatsinski's (2017) participants tended to extend frequent forms to new meanings in production but tended to map them onto the experienced meanings in comprehension. When the same form was rare, it was extended to a new meaning less in production, and was mapped onto it more in comprehension. Thus, frequent forms are extended to new meanings because they are more accessible than rarer forms, *even when* the less frequent form would be a better expression of that meaning (see also Koranda, Zettersten and MacDonald 2022, where match to the meaning is quantified objectively).

Even though a rare form may often be worth waiting for, the speaker who already accessed the frequent form and is deciding whether or not to plan more or start speaking (dashed line in Figure 1) has no way of knowing whether a better alternative to the form they have accessed will eventually come to mind. (The contestant does not have the form *lie* accessible when they decide not to say *lice* – it takes him another minute to come up with the form.) Therefore, for the speaker to delay speaking, they must think that there is something wrong with the form they have accessed. That is, the speaker cannot wait to suppress producing the frequent form by comparing how well it expresses the meaning compared to an infrequent form. The suppression must often not be driven by a comparison of alternative expressions but by an evaluation of the dominant expression.

Kapatsinski (2022) proposed a processing mechanism by which a form may be detected to be unsatisfactory before any other form is accessed, allowing the speaker to delay production. This mechanism is illustrated in Figures 2–3. Follow-

ing the connectionist framework, the mechanism assumes that processing works via spreading activation and inhibition.

In Figure 2, the accessed form sends feedback to semantics. Like in comprehension, the amount of feedback reaching a meaning from a form is proportional to how well the form cues the meaning; p(meaning|form) or Δp = p(meaning|form) – p(meaning|–form) depending on learning model (e.g., Gries and Ellis 2015; Kapatsinski 2018; Kapatsinski and Harmon 2017; Ramscar, Dye and Klein 2013). However, unlike in comprehension, this feedback – localized within the production system – is inhibitory: it reduces the activations of meanings that are strongly cued by the activated form.

For meanings that are part of the intended message (LOUSE in Figure 2), this negative feedback makes little difference because they are receiving strong excitatory input from the message. However, any meanings cued by the form that are not part of the intended message now have a negative activation level (i.e., inhibition).

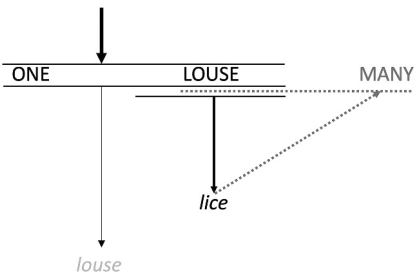

Figure 2: The form *lice* inhibits the meaning(s) it cues (shown by the red dashed arrow). The LOUSE part of the meaning remains activated because it is still receiving excitation from the message. But MANY is now inhibited and has some inhibition to pass on.

The inhibition then spreads from unintended semantics back down to the associated form(s), inhibiting them (Figure 3). Because feedback inhibition cycles back down to the form(s) that generated it, this mechanism is called the Negative Feedback Cycle (NFC). As a result forms that would strongly activate unintended meanings in comprehension are inhibited, and the speaker continues planning (Figure 3). Thus, the NFC allows the speaker to avoid producing frequent forms when they are likely to have unintended consequences.

4 Side Effects of the Negative Feedback Cycle

The primary function of the NFC is to prevent blurting out overextensions, which look creative but – on the production-internal view of creativity – aren't. However, despite its role in enforcing convention, the NFC can occasionally produce behaviors that both look and are creative.

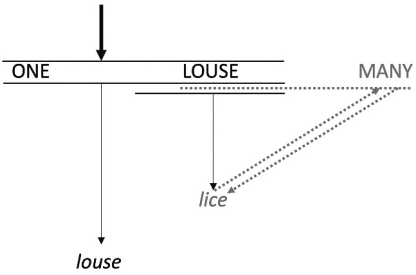

Figure 3: Inhibition then spreads from the unintended meaning MANY back down to the activated form(s) that cued it (*lice*), deactivating or inhibiting them. This both buys the speaker time to activate the initially less accessible form *louse* and ensures that it wins the competition against the initially more accessible form (*lice*).

The NFC suppresses the production activated first, or most strongly by the context when it is estimated by the speaker to be likely to have unintended consequences – to be misinterpreted by the listener as expressing a message that the speaker does not intend. Suppression of this production results in the selection and execution of a production that is less likely given the context. These productions are creative in the production-internal sense – they are not the most expected productions given the context, and require the speaker not to blurt out the first thing that comes to mind. They are also relatively effortful and take some time to produce – inhibition needs time to cycle. They can also look highly creative to an observer once produced. Nonetheless, they too are the product of the normal functioning of the production system.

4.1 Subtraction

The first creative consequence of NFC is deletion of units that express unintended meanings from larger forms, when no smaller form to express the intended message is available. The clearest example of such creative deletions is backformation, which refers to a process by which a speaker generates a new form by deleting what looks like a morph from a pre-existing form. For example, the speakers who first produced *edit* from *editor*, *burgle* from *burglar*, *destruct* from *destruction*, or *pea* from *peas* engaged in backformation. Currently these productions are of course no longer creative.

Let us look more closely at the case of *editor*. The speaker who created the verb *edit* must have wanted to express the message "perform the job of an editor" or "do what editors do", the act of editing. In the absence of the word *edit*, the closest expression of this message EDIT_{ACT} would be the word *editor*. The speaker had the option of just verbing it: after all, we *author* papers and *engineer* language models rather than *authing* and *engining*. Simple conversion of nouns into

verbs, verbing, is by far the dominant way of forming verbs in English. However, the speaker decided to delete *-or*. Why? Presumably because *-or* has unintended semantics – it is a very good cue to agentivity, ONE.WHO.[. . .]$_{ACTS}$, and this is not part of the intended message.

More formally, we can describe the process as in Figure 4. The speaker's message EDIT$_{ACT}$ first activates the closest matching form, *editor*, as there is no form *edit* yet. At this point, *editor* could have been converted into the verb *editor* and produced to mean EDIT. However, the *-or* is a good cue to ONE.WHO.[. . .]$_{ACTS}$ = AGENT. It therefore inhibits these unintended semantics, which inhibit it in return through the Negative Feedback Cycle. A similar process can account for other instances of backformation in which a morph is deleted, like *peas* > *pea*, where *-s* would be inhibited by the unintended PLURAL meaning.

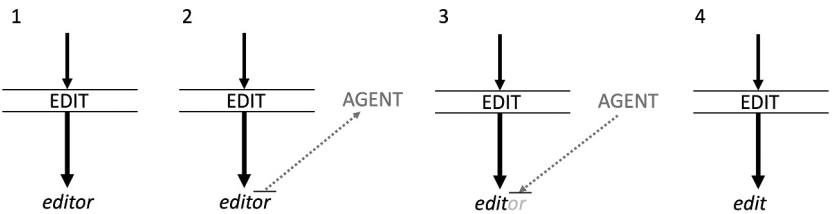

Figure 4: Left: First stage: *editor* activated by EDIT. Middle: Second stage: *-or* inhibits AGENT and is inhibited by it. Right: *-or* is inhibited by inhibition cycling back from AGENT and *edit* is produced.

Notice again that there are verbs like *author* and *engineer*. Backformation is a sporadic process. This sporadicity is actually expected under the NFC account: the NFC needs time to act, and backformations should therefore take time for reflection. After all, four out of five adults asked for the singular form of *lice* in our initial example, quickly wrote *lice* rather than backforming *lie*. Most of the time the speaker does not have enough motivation to wait (production errors do not usually cost thousands of dollars), and may have motivation to start speaking as quickly as possible (Holler et al. 2021). An interesting prediction of the NFC account is therefore that backformations can be distinguished from speech errors by being less, rather than more, likely to occur under time pressure.

4.2 Avoidance and Circumlocution

The strongest evidence for NFC is provided by avoidance behaviors, where the speaker avoids a form that is demonstrably the most likely one given the intended message. Avoidance results in selection of a less likely form, a (possibly

novel) circumlocution, or sometimes nothing at all (i.e., a paradigm gap). In either case, the suppression of the most accessible form is a necessary prerequisite for the emergence of a novel creative solution.

Avoidance is strongest and most successful when the avoided form has taboo connotations. Specifically, Motley et al. (1982) and Dhooge and Hartsuiker (2011) show that speech errors that would result in taboo utterances (like the exchange of initial consonants in *hit shed*) are avoided more successfully than errors that would not result in a taboo utterance (e.g., a similar exchange in *hip shed* would be more likely to be produced). Dhooge and Hartsuiker (2011) further show that speakers take longer to initiate word production when a taboo utterance is likely to result from an error. These results suggest suppression of taboo words like *shit* before they are executed – something the NFC is designed to account for.

Trask (1996) provides several textbook examples of lexical replacement due to the original word becoming tabooed. One well-known example is the replacement of the Proto-Indo-European word bear by *m^{j}edv^{j}ed^{j}* < *med-o-jed* 'honey eater' in Russian (Fasmer 1986).[6] Another is the avoidance of *lie* in the sense of lying flat (rather than being untruthful) in favor of *lay* as in *I would lay on the couch* (COCA).

The NFC provides an account of these types of replacements, as shown in Figure 5. In the first diachronic stage of the language, *bear* is the normal way to say BEAR. However, when *bear* becomes tabooed (for whatever social reason) it acquires additional connotations. That is, the listener would think "I can't believe he just said *bear*!" (or *shit* or *God* or the name of a dead relative depending on the particular nature of the taboo). The negative nature of the connotation means that it has negative activation by default, and this negative activation can always spread to corresponding forms. In fact, the forms themselves will be likely to have a negative resting activation level as a result.

The negative resting activation can be overcome by excitation coming from the message – i.e., when the connotation is intended. Harry Potter can say *Voldemort*, and Voldemort can say *avada kedavra*, even though both utterances are heavily tabooed for wizards in the universe. Harry and Voldemort say these things because they intend the tabooed consequences (Voldemort intends killing someone by saying *avada kedavra*). However, when a tabooed meaning is unintended, its default negative activation level spreads to the corresponding forms

6 Following Fasmer (1986), I assume that the form was originally *m^{j}ed-o-jed* 'honey eater', with the common compound interfix *-o-*, [o] reduced into a glide in this common form, and the form was then reinterpreted as 'honey knower'. However, it could also be assumed, following other etymologies, that *ved* 'know' is the original form.

and makes them particularly easy to inhibit and therefore avoid producing. (If the taboo is strong, one really needs to make an effort to say the tabooed form even when it is intended, keeping the message in mind for longer until the negative resting activation is overcome.)

Returning to our bear example, as *bear* is suppressed, less likely forms can win the competition. *Honey eater* is one of the many possible such forms that can become conventionalized. Its initial production by some past speaker seems undeniably creative, as *bear* would be a more accessible form at the time, and would probably have seemed creative at the time (imagine someone referring to bears as *honey-eaters* in English!). As illustrated in Figure 5, this production likely comes from the semantics of BEAR containing properties like eating honey, which activate associated forms.

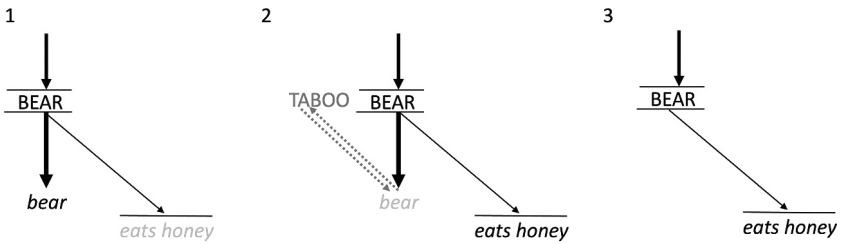

Figure 5: Left: Initially, BEAR strongly activates *bear* and weakly activates semantically similar utterances one could use as circumlocutions like [that creature that] *eats honey*. Middle: *bear* cues the taboo connotation and is therefore inhibited. Right: This leaves *eats* and *honey*, which are then slotted into the common $[. \ . \ .]_N$-o-$[. \ . \ .]_N$ construction for nouns referring to agents of transitive actions (the unification with the construction not shown; see Dell 1986; Kapatsinski 2017, 2021 for possible mechanisms).

Another instructive example discussed by Trask (1996: 41) is avoidance of forms that resemble names of dead relatives or community members. For example, the death of a community member called *djajila* in 1975 led speakers of the same community to avoid the verb *djäl*, which until then was the most common way to say WANT. The verb was replaced by the hitherto less frequent alternative *duktuk*. This example is interesting because it is not only the form *djajila* that is avoided: forms that are similar enough to *djajila* to activate its meaning are avoided as well. Thus, even though *djäl* means WANT, its production is suppressed because it activates DJAJILA, and the associated memories, enough. The phenomenon is illustrated in Figure 6.

This example supports the NFC thesis that taboo avoidance comes about because the speaker notices, implicitly, that the form they are about to produce would have unintended consequences. In other words, the form selected for pro-

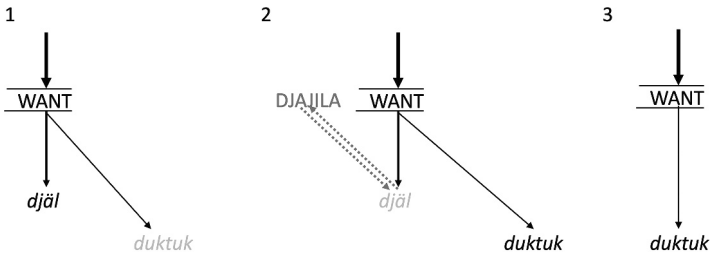

Figure 6: Left: Initially, WANT strongly activates *djäl* and weakly activates *duktuk*. Middle: *djäl* cues the semantics of a dead community member DJAJILA and is therefore inhibited. Right: This leaves *duktuk* to win the competition for WANT.

duction is the ultimate source of inhibition. A speaker who is about to say *djäl* would not have DJAJILA activated as part of their message: WANT and DJAJILA are semantically unrelated, so the message is unlikely to contain, or even activate DJAJILA. The only reason DJAJILA would come to mind is the accidental resemblance between the form *djäl* and the form *djajila*. Access to *djäl* is therefore necessary to activate the taboo semantics, launching the NFC. Furthermore, this example supports the associative nature of the NFC: the NFC suppresses not only forms that *refer* to a taboo meaning, but forms that strongly *cue* a taboo meaning. The form *Djajila* is suppressed most strongly only because it is the best cue to DJAJILA: forms that merely evoke DJAJILA can also be suppressed. In this case, suppression of the regular, run-of-the-mill way of expressing the meaning WANT leads the speaker to produce the less likely *duktuk*, initially a borrowing from another language. This may be akin to an English speaker replacing *want* with *desire*, which would likely be perceived as creative. In *Lord of the Rings*, Celeborn saying *Where is Gandalf, for I much desire to speak to him* is one of the most memorable lines in the story, and works well in conveying the ancient and ethereal character of the elves, in part because of the unusual verb *desire*. However, whether or not it is *perceived* as creative, the use of *duktuk* in place of *djäl* is creative for the speaker – it requires suppressing the usual way of expressing a meaning after it has come to mind enough to activate the taboo semantics.

Another consequence of the NFC is Gresham's Law of Semantic Change – "bad meanings drive out good" (on analogy with the original Gresham's Law, "bad money drives out good", in economics; Burridge 2012; Trask 2003: 45). By providing a mechanistic account of Gresham's Law, the NFC accounts for the common semantic change of pejoration. Specifically, pejoration occurs as a sequence of two changes: extension to a new but related meaning, which just happens to be negative or tabooed, followed by avoidance of the term when the new meaning is not intended. For example, consider the word *intercourse*. The Corpus of Histori-

cal American English (COHA, Davies 2012) shows numerous 19th century examples of its use to mean EXCHANGE or INTERACTION. For example, Jane Austen in *Pride and Prejudice* writes that *Mr. Darcy and Elisabeth had no intercourse but what the commonest civility required*, by which she means that they barely exchanged a word. There are also numerous 19th century bureaucratic documents with titles like *Rules and regulations concerning commercial intercourse with and in states and parts of states declared in insurrection* (from 1864, the American Civil War), where the word means an exchange (of goods). However, around 1890, *sexual intercourse* begins to appear. This of course is a simple extension to a new context (SEXUAL INTERACTION). However, *intercourse* is now a cue to SEX. Since SEX is a taboo meaning, the NFC will now suppress the production of *intercourse* when SEX is not intended (as part of the message; Figure 7). This means that *intercourse* stops being used in non-sexual contexts and therefore strengthens its co-occurrence and association with SEX. As a result, *intercourse* can now mean SEX without the word *sexual*. As *intercourse* can no longer be used to mean INTERACTION, other words for the same concept must take its place – the speaker needs to come up with some solution to this problem. Thus, the word *interaction*, previously rare, is increasingly selected for production when SEX is not intended, and rises in frequency (about 10-fold from 1890 to 1980 in COHA). This would be quite akin to a current speaker avoiding the term *interaction* by referring to a verbal interaction as a *thought-dance,* which I have obtained from ChatGPT by telling it to be creative instead of giving me established synonyms for *interaction* (prompts: *synonyms for interaction* followed by *Can you make a creative one?*). By avoiding the common terms, it is possible to have the same concept generate a less common, or even novel form.

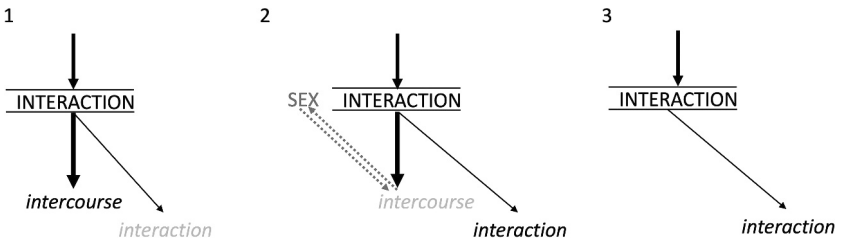

Figure 7: Left: Initially, INTERACTION strongly activates *intercourse* and weakly activates *interaction*. Middle: after 1890, *Intercourse* cues the semantics of SEX is therefore inhibited. Right: This leaves *interaction* to win the competition for non-sexual INTERACTION. Conversely, because *intercourse* now occurs only when SEX is intended, the association between *intercourse* and SEX continues to strengthen.

Not all avoidance behaviors are driven by taboo. Sometimes, the high likelihood of misinterpretation is sufficient. For example, Ceuppens and DeSmet (2024) show that when a form is extended to a new meaning by inference-driven metonymy, the original meaning of the form is often lost, but when the extension is metaphorical, the original meaning persists. They argue that metonymic extensions tend to result in an ambiguous form whose ambiguity is not resolved by context. Indeed, the existence of bridging contexts in which the form is ambiguous between old and new meanings is what allows inference-driven metonymy to occur (Traugott 1988). For example, in contexts like *It's been disaster after disaster since Clark was elected president*, *since* can be interpreted as both AFTER and BECAUSE, allowing for the metonymic extension of *since* from AFTER to BECAUSE. As a result, one might expect avoidance of *since* in contexts where BECAUSE is not intended (and this is observed by Ceuppens and DeSmet). In contrast, metaphoric extensions do not require ambiguous contexts, and tend to result in ambiguity that is resolved by context. For example, one knows whether one means a literal chair or the chair of some organization when the word *chair* is used. As a result, the use of the form in the original meaning is not avoided, allowing that meaning to survive.

Misinterpretation-driven avoidance is controversial in morphology, but is now well documented in phonetics. For example, several researchers have shown that speakers hyperarticulate the phonetic cues that have just been misperceived by the interlocutor (e.g., Buz, Tanenhaus and Jaeger 2016; Schertz 2013; Wedel, Nelson and Sharp 2018). For example, speakers increase average voice onset time (VOT) in *cod* if it the listener misperceived as *God* last time it was said, but not if it was misperceived as *pod*. Importantly, the hyperarticulation in question appears to be based on avoidance of the most ambiguous articulations (here, short, [g]-like VOTs), rather than targeting of the least ambiguous articulations (super-long VOTs). Buz and colleagues find that misperception-driven hyperarticulation results in tucking in of the left tail of the distribution of voiceless VOTs, rather than extension of the right tail or a shift in the central tendency. While there is no space to address this phenomenon fully here, this is what one would expect from the NFC, because what the NFC suppresses most strongly are potential productions that are most likely to be misperceived (here, productions with short VOT). Indeed, Stern and Shaw's (2023) successful model of misperception-based hyperarticulation implements the same principle, in suggesting that hyperarticulation involves inhibiting potential productions in proportion to the frequency with which they realize the unintended phone or message.

The NFC provides a novel perspective on the role of homophony avoidance in the emergence of paradigm gaps. A gap is an environment where creativity is required of the speaker to express the message because all conventional expres-

sions of the message have been suppressed. For example, in Spanish, a famous paradigm gap in the first person singular present is *abuelo* (Figure 8), whose production as the first person singular of the verb *abolir* 'abolish' is avoided. The NFC suggests, contra most other accounts of gaps (Albright 2003; Gorman & Yang 2019; Sims 2015), that it is not an accident that *abuelo* is the word for GRANDFATHER in Spanish. Crucially, the meaning GRANDFATHER is far more frequent than ABOLISH. Therefore, when the message I.ABOLISH activates the form *abuelo*, the form will cue GRANDFATHER much more strongly than it cues I.ABOLISH. As a result, its production is likely to be suppressed. Because there is no other active alternative, a paradigm gap results, as shown in Figure 8.

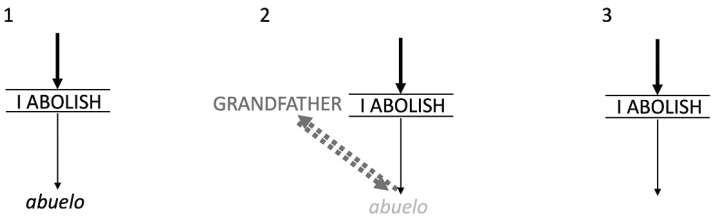

Figure 8: Left: I ABOLISH activates *abuelo*. Middle: *abuelo* strongly cues GRANDFATHER and is therefore suppressed. Right: nothing is left to say.

Many researchers are skeptical of homophony avoidance as an explanation for this gap. One objection is that avoidance occurs in other paradigm cells where there is no complete homophony. Albright (2003: 8) writes "not all parts of the paradigm would be affected by homophony, so even if *abuelo* happens to mean 'grandfather', there would be no reason to avoid the 3pl *abuelen*, which is not a possible noun form". However, complete form overlap is not necessary for avoidance, as exemplified above by the avoidance of *djäl* because it is close enough to activate DJAJILA. It is sufficient for the form to cue an unintended meaning. The sequence *abuel . . .* is surely more than enough to cue GRANDFATHER, which is much more likely than ABOLISH. Far shorter and more ambiguous word-initial chunks have been shown to elicit activation of meanings of the most likely words, and even their semantic associates. For example, in visual world eyetracking studies, listeners look at referents of words that begin with what they have heard so far (Allopenna, Magnuson and Tanenhaus 1998; Teruya and Kapatsinski 2019), and even at their semantic associates (Yee and Sedivy 2006) more than they look at pictures of unrelated meanings. Similarly, Pirog Revill et al. (2008) show, using fMRI, that words with no motion semantics activate the motion area of the brain (MT) if they overlap phonologically with motion words by one syllable. Thus,

abuel is likely enough to activate the semantics of *abuelo* and result in NFC suppressing the activated form.

Another objection is that there are other forms that have homophones and are nonetheless produced (Albright 2003; Gorman and Yang 2019; Halle 1973; Sims 2015). For example, Albright (2003: 8) writes "most importantly, there are many cases in which homophony is tolerated: *creo* 'I create'/'I believe', *avengo* 'I avenge'/'I reconcile', *suelo* 'I am used to'/'I pave', etc". However, none of these cases have the massive imbalance in token frequency between the intended meaning and the unintended meaning that is true of *abuelo*. For example, in the Corpus del Español (corpusdelespanol.org, Davies 2002) there are 79 *abolir* 'abolish' and 1266 *abuelo* 'grandfather', compared to 2894 *creer* vs. 1941 *crear*.[7] The size of the token frequency asymmetry is predicted to be crucial for the avoidance to happen by the NFC: *abuelo* is a much stronger cue to the unintended meaning (GRANDFATHER) than *creo* is.

Consider also the following example from Russian (Figure 9; raised by Halle 1973, and echoed in both Gorman & Yang 2019, and Sims 2015, despite their theoretical disagreements). Russian has gaps in the 1st person singular non-past in verbs of the *-i-* conjugation, in which stem-final coronals become alveopalatals, e.g., [d] becomes [ž]. For example, *deržu* is the expected 1st person singular non-past of the verb *derzit^{j}* 'to speak impudently'. The NFC suggests that *deržu* is avoided when the speaker tries to produce I.SPEAK.IMPUDENTLY, a rare expression, because *deržu* is also the 1st person singular non-past form of a far more frequent verb, *deržat^{j}* 'to hold'. According to the Russian National Corpus (available at ruscorpora.ru; see also Grishina 2006), *deržat^{j}* is 300 times more frequent than *derzit^{j}* (77562 vs. 252; lemma search in the main corpus on 8/27/23). Therefore, if produced, *deržat^{j}* would cue the unintended message I.HOLD more strongly than the intended message I.SPEAK.IMPUDENTLY. It would therefore be suppressed by the NFC.

An objection to this reasoning is that ambiguity is tolerated in the form *vožu*, which could mean either I.DRIVE.VEHICLE/I.LEAD.AROUND (*vodit^{j}*) or I.CARRY.BY.VEHICLE (*vozit^{j}*). However, searching the Russian National Corpus reveals that the two verbs are about equally frequent: both are about 10 times the frequency of I.SPEAK.IMPUDENTLY (11328 vs. 9985 respectively). Therefore, inhibition from the unintended meaning would be counterbalanced by activation from the equally frequent intended meaning. The NFC is therefore less likely to succeed in

7 The other two examples where ambiguity is supposedly tolerated are not findable: *avenir* has a single token, while no examples of *solir* or *soler* are found.

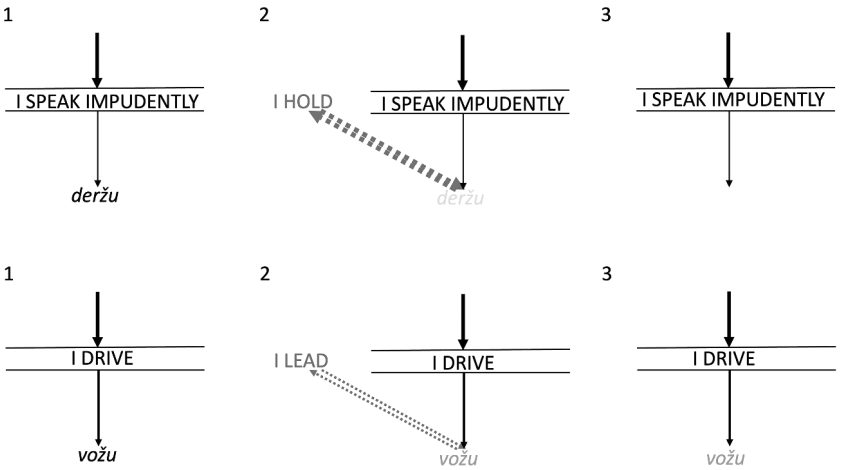

Figure 9: Top: *deržu* strongly cues the unintended message I.HOLD and is suppressed by it. Bottom: *vožu* does not strongly cue the unintended message I.LEAD and survives.

suppressing *vožu* in either sense than *deržu* in the sense of impudent speaking, explaining why only the latter is avoided. This difference is illustrated in Figure 9.

Finally, another objection to ambiguity avoidance as a source of gaps is that there are gaps not explained by this mechanism (e.g., Albright 2003: 8, mentions two verbs in Spanish that have gaps despite absence of homophony). However, we do not expect all gaps to arise for the same reason: there are many reasons that a form might be unacceptable. For example, Russian feminine nouns that have no vowel in the stem have gaps in the plural genitive because deletion of the suffix would leave them with no vowels, e.g., *xna* would become *xn*. Although gaps like this can also be explained by avoidance, it would be avoidance caused by pronunciation difficulty (see also Berg 1998; Martin 2007; Schwartz & Leonard 1982; for other cases of such avoidance). This kind of avoidance could be caused by negative feedback to form selection, but from articulation rather than from semantics (Berg 1998; Martin 2007) or through experience (trying to say [xn] and not liking the consequences; Kapatsinski 2018). Other forms might be gapped because unacceptability becomes associated with certain sublexical chunks through generalization from forms that do have infelicities that cause them to be avoided or stigmatized (Daland, Sims and Pierrehumbert 2007).

Our last example illustrates that avoidance is caused by ambiguity only if one of the meanings is unintended. In some cases, the form is intended to bring to mind another referent. Poetry of course comes to mind – a good poem should have more than one interpretation – but a more prosaic example is presented by

names in societies where children are often named in honor of their parents, other relatives or people the namers admire (e.g., *Albus Severus Potter*). For an accessible example, my name *Vsevolod* was given to be in honor of my great-grandfather and was intended to bring him to mind. Although this is an extension of a form to a new referent, it is a deliberate one, and not entirely accessibility-driven as the name is otherwise rare, and the naming occurred many years after my greatgrandfather has died.

Hypocoristics bring out the interplay of ambiguity avoidance and ambiguity-seeking well. For example, all Russian names have conventional hypocoristics. There are two conventional shortenings for *Vsevolod*, *Seva* and *Vol^{j}a*. The former is more common, and yet, *Vol^{j}a* was selected for me because it matched my great-grandfather, who was intended to come to mind when the name is uttered. In contrast, unintended ambiguity is avoided. For example, one would think that *Volod^{j}a* is a possible shortening of *Vsevolod*. However, it is a conventional shortening of the much more frequent name *Vladimir*. As a result, its use to refer to *Vsevolod*s is blocked (in fact, I am frequently called *Volod^{j}a* by non-native speakers, who overextend the name). (While *Seva* might be argued to be favored over *Volod^{j}a* because it preserves the stressed syllable, this is not true of *Vol^{j}a*.)

As in the case of gaps, ambiguity in hypocoristics appears to be tolerated when frequency asymmetries are smaller: *Slava* could be the shortening to a wide range of names ending in *slav*: *Vladislav*, *Rostislav*, *Izjaslav*, *Svjatoslav*. It is probably not an accident that these names are relatively uncommon: *Vsevolod* vs. *Vladimir* shows a huge frequency asymmetry (2700 vs 32000 in the Russian National Corpus)[8] while the *slav*s are more closely matched in frequency (1600, 1500, 800, 600, respectively. It is therefore likely that *Slava* will not activate the wrong Slava in context, while *Volod^{j}a* naming a Vsevolod is likely to. The overall lower frequency of the *Slava* names is also relevant: one is likely to know a *Volod^{j}a* who is a Vladimir when naming a Vsevolod but is less likely to know a *Slava* who is a Vladislav when naming a Rostislav. Finally, individuals avoid naming children after people they dislike, and one is more likely to dislike a Vladimir than a Rostislav, simply because there are so many more Vladimirs.[9] Therefore, the unin-

8 Here and subsequently, Russian frequencies were obtained by using wordform search for the base Nominative Singular form in the main subcorpus of the Russian National Corpus (Savčuk, Arxangelskij, Bonč-Osmolovskaja, Donina, Kuznecova, Ljashevskaja, Orexov and Podrjadčikova 2024) and rounding to the nearest 100.

9 These frequency asymmetries might be somewhat skewed by Putin, who is a *Vladimir*, but should hold nonetheless: Vladimirs are numerous, so there will often be at least one politician named *Vladimir* (e.g., Lenin and Zhirinovskij were Vladimirs as well) causing its frequency to skyrocket in Zipfian fashion.

tended connotations of the former should be much more effective in driving the NFC. The contrast is illustrated in Figure 10.

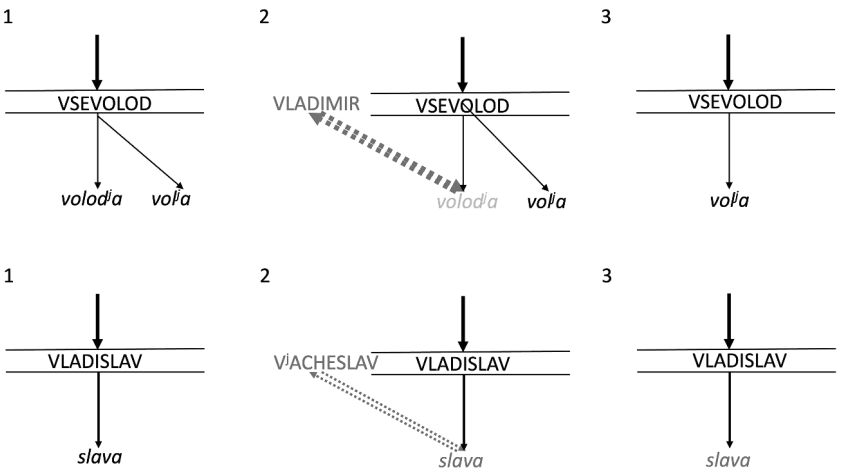

Figure 10: Top: *volodʲa* strongly cues the unintended referent VLADIMIR and is suppressed by it in naming VSEVOLOD. Bottom: *slava* does not strongly cue the unintended referent VʲACHESLAV and survives in naming VLADISLAV.

The need for creativity often arises from ambiguity avoidance in naming when one becomes friends with two people who have the same full names and hypocoristics (a common occurrence). In such cases, people will often deliberately change the hypocoristic of one of the people because it evokes the other, or devise ad hoc nicknames to differentiate them.[10]

10 Note that Stankiewicz (1957: 199) proposed that ambiguity is common within gender but avoided between genders. I would tentatively suggest the opposite, taking into account frequency asymmetries: in the case of different-gender names that differ only in the suffix, like *Valentin/Valentina, Aleksandr/Aleksandra*, ambiguity is largely tolerated; perhaps because the frequency differences are not dramatic but also because the gender of the referent and the agreement markers on verbs and adjectives tend to disambiguate the referent in context. So *Stepanida* can be *Stëpa* despite the greater frequency of *Stepan* (11000/1600). For within-gender names, the pairs in Stankiewicz (1957) are of names of similar frequency mentioned above, and very rare names that are obsolete, making the preferred hypocoristics difficult to verify (*Mitrofan*, *Valerian*), e.g., were Valerians mostly Valʲas like Valentins or Valeras like Valerijs?

The most problematic example is *Mitja* as the shortening of both *Dmitrij* and *Mitrofan*. *Dmitrij* is much more frequent than *Mitrofan* (15.4K/0.9K) and *Mitʲa* is slightly more common than *Dima* (9.6K *Mitja* vs. 7.2K *Dima*), an alternative shortening for *Dmitrij* stated to be rare in Stankiewicz (1957). Both hypocoristics are also much more frequent than *Mitrofan*. It is possible that *Mitrofan*

5 General Discussion

The Negative Feedback Cycle suppresses productions that are likely to have unintended consequences; more specifically, productions that are likely to transmit unintended meanings. Importantly, suppression is only needed when the risky production is also the most likely production given the context – the first expression that comes to mind. Because of this, the NFC results in productions that are not only safer, but also often perceived as creative – circumlocutions to fill a paradigm gap, backformations, novel or unusual hypochoristics, and generally forms that the speaker would not have chosen to express the same meaning without a moment of reflection.

It appears impossible to avoid postulating something like the NFC if one takes seriously the finding that forms can be produced even when they do not fully match the speaker's intended message – simply because they are more accessible than forms that would express the message better (V. Ferreira & Griffin 2002; Harmon & Kapatsinski 2017; Koranda, Zettersten and MacDonald 2022), and yet speakers have the choice to continue planning and avoid blurting out the first thing that comes to mind. In the extreme context of a talk show, the example we started with, when ten thousand dollars are on the line, the speaker can spend several minutes trying to come up with the right word. To do this, the speaker must suppress production of forms that have already been activated as *not good enough to produce*. The idea of good-enough production was recently emphasized by Goldberg and F. Ferreira (2022) and Koranda et al. (2022). However, good-enough production begs the question of how the speaker would know whether what they are about to say is good enough. The NFC provides the first explicit mechanism by which the speaker could accomplish this goal. The fact that the NFC also accounts for a number of creative behaviors in language production, and makes novel predictions about these behaviors (such as the role of frequency asymmetries in gaps) is a pleasant side effect. Nonetheless, all of the behaviors discussed here demand proper studies that cannot be accomplished here in the available time and space, but represent promising directions for future work.

---

was only ever *Mit^{j}a* (since usually the first syllable is retained in a hypocoristic) and *Dmitrij* also strongly tended to be *Mit^{j}a* at the same time. This would suggest parents of Mitrofans tolerated the ambiguity with *Dmitrij*. However, this is hard to test in the corpus because *Mitrofan* de facto has no hypocoristics: of the first 100 examples of *Mitrofan* in the Russian National Corpus, all are names adopted by adults when they became monks in the Orthodox Church, abandoning their prior secular name, and monks are not referred to with hypocoristics.

The demise of these names whose likely hypocoristics are likely to be misinterpreted may not be an accident. In other words, instead of developing a new creative hypocoristic for a rare name to avoid misinterpretation, one could also avoid the full name, making it even rarer.

In particular, the NFC makes specific predictions about when a form's production is likely to be suppressed, given sufficient time: when the form has a taboo meaning (and that meaning is not intended), when the form has many specific but unintended meanings, and when the unintended meaning(s) of an ambiguous form are frequent relative to the intended meaning. All of these characteristics of a form, importantly, are expected to matter specifically when the NFC has had time to operate: early in processing, the probability of a form's production should depend only on the degree to which it is cued by the message – increasing with the number of semantic features of the form activated by the message times the probability of the form given each feature, and decreasing with the number of the form's semantic features inhibited by the message (Kapatsinski 2022).

The specific characteristics of the NFC are also, of course, up for further investigation and debate (see, e.g., Chuang et al. 2021; Dhooge and Hartsuiker 2011; Hartsuiker and Kolk 2001; Nozari, Dell and Schwartz 2011, for related ideas about how monitoring and suppression might work). The present paper has only scratched the surface of the field by digging up a few illustrative examples. For example, I have assumed that there must be something wrong with the form that a speaker is about to produce for them to reject it – the accessed form inhibits itself because it is a cue to unintended semantics. This is clear in the example of taboo avoidance driven by phonological similarity to the taboo form, in the absence of any semantic similarity.

Alternatively, a careful speaker may delay execution regardless of the appropriateness of what they have planned, and continue pumping activation into the system from the message until all possible alternatives are activated. At this point, the speaker may be able to compare them on how well each alternative production would express their intended message, with the eventually selected form matching the message better because it does not activate any unintended semantics. The NFC is only preferred over this alternative on a priori grounds at present: it has the functional advantage of delaying production only when a delay is needed, and is computationally simpler because it does not require comparison operations. Comparing the meaning activated by each form and the intended meaning would require computing a predicted semantic vector for each form in working memory to compare with the intended semantic vector. However, this advantage in prior probability could be overturned by empirical findings showing that the appropriateness of the original form accessed has no effect on how long a careful speaker takes to plan an utterance. This could be investigated, for example, by priming contextually appropriate vs. inappropriate forms along the lines of V. Ferreira and Griffin (2002).

The NFC proposes that the speaker decides to avoid starting to speak before having accessed an appropriate replacement for the initially accessed form. Alternatively, one could propose that the form eventually produced is what blocks the production of a more frequent, primed, or otherwise accessible form (a mechanism referred to as blocking in Aronoff, 1976, or statistical pre-emption in Boyd and Goldberg, 2011). However, blocking and statistical preemption do not account for the existence of defectivity / paradigm gaps, where the speaker struggles to come up with *any* acceptable production for a while. To return to our initial example of a talk show contestant struggling to produce *lie* as the singular form of *lice*, the contestant does not know what to say for several minutes, but still avoids producing the only form of the word that he does know.

The NFC proposes that the activated form sends *inhibition* up to the semantics it cues. This would, of course, be a non-starter if the feedback took place in the comprehension system. In comprehension, the form *activates* the unintended semantics, rather than inhibiting them. However, there is good evidence that feedback in production is internal to the production system and separate from the comprehension system. In particular, error monitoring (which is function of the NFC) appears to be production-internal because it can be damaged in aphasia independently of comprehension (Hartsuiker and Kolk 2001; Nozari, Dell and Schwartz 2011). The main motivation for the bottom-up inhibition is implementational simplicity – by assuming that the form sends up inhibition, the NFC can be implemented using exclusively spreading inhibition. If the form were sending up excitation, we'd have to somehow turn it into inhibition before it comes back down. However, this is again an empirical hypothesis. The present proposal suggests that unintended meanings associated with the suppressed form should be inhibited, and therefore harder to activate in the immediate future.

The NFC takes time to operate. As a result, when the speaker needs to start speaking quickly (e.g., in a multi-party conversation where other speakers would jump in at any sign of hesitation, Holler et al. 2021), the NFC may not have time to suppress accessibility-driven production choices. Conversely, a writer of a research article like this one – who has nearly unlimited time to plan, and Reviewer 2 to contend with – will often produce and discard multiple possible formulations of the same message because all end up having unintended interpretations, and the consequences of misinterpretation are relatively severe.

One interesting question at the form level is what is suppressed when a writer decides that an abbreviation is not sufficiently unambiguous. For example, in taking notes in the margins of a book, I recently initially wrote down *habit* as an abbreviation for *habituation* but, realizing that I would be likely to misinterpret *habit* as HABIT, continued into *uat*, after a moment's hesitation. Although this is a case in which the producer continues producing, rather than continuing

planning, upon reflection, it would be desirable to account for this phenomenon with the same mechanism as the cases of ambiguity avoidance we have discussed. However, what is being suppressed here? An interesting possibility is that what is suppressed is the action of stopping production (what Diesburg and Wessel 2022 call the "cancel process"), rather than the production *habit*. However, NFC would not be able to suppress it because the action of stopping is not associated with the meaning HABIT. From the NFC perspective, we are forced to assume that what is suppressed is *habit*, allowing the otherwise more costly *habituat* to win. A possible advantage of this account is that it explains why typing was not stopped after *u* or *a*, where the string is as unambiguous as after the second *t*: *habituat* is a chunk (stem) while *habitu* and *habitua* are not.

At the semantic level, one question is whether discrete semantic nodes are needed, or if semantics can be represented as a continuous space (e.g., Chuang et al. 2021). The phrasing of the present paper suggests that semantic representations are composed of discrete unary features like PLURAL or BOVINE. With this representational format, hypernyms are special: *thing*, *stuff*, and *this* may not be effectively suppressed by NFC in producing more specific words because they do not have any unintended semantic features. This may not be desirable because speakers are sometimes dissatisfied with a hypernym and produce a hyponym to it upon reflection (e.g., replacing *dog* with the name of a particular breed). If so, then the absence of a feature could be an unintended consequence, and unary features would be insufficient. Ultimately, points, regions, or distributions over semantic space may also present a better alternative to features (e.g., continuous patterns of activation over a set of hidden nodes as in Rogers and McClelland 2004; or dynamic neural fields, as in Stern and Shaw 2023).

Returning to the nature of creativity, the connectionist framework within which the NFC account is situated provides a different perspective on creativity than traditional symbolic grammar. Specifically, Sampson (2016) has argued that the simple application of existing constructions (form-meaning mappings) to new input forms (i.e., productivity) is distinct from creativity, or F(ixed)-creativity vs. E(nlarging)-creativity in his terms. (E-)creativity requires extension beyond the system, breaking the rules. While this distinction is intuitive, it presupposes that linguistic generalizations rely on classical categories where an input either is or is not eligible to undergo a particular rule (see also Hoffman 2019). From a connectionist perspective, forms do not have necessary and sufficient conditions on use; the selection of a form depends on simultaneous combination of a multitude of contextual and semantic influences (e.g., Kapatsinski 2009). In such a system, extension is an inevitable side effect of the distributed nature of mental representations and cannot be distinguished from following the rules (Bybee & McClelland 2005; see also Suttle and Goldberg 2011, for a related perspective). Extensions can

vary in how similar the original use of a form is to its new use, and in how they are perceived by listeners, but all rely on the same basic mechanism – activation of forms by distributed semantic patterns. From this perspective then, extensions – no matter how creative-looking – are not true (E-)creativity because the producer simply says the first thing that comes to mind in accordance with the normal functioning of the system. Creativity requires following the path less traveled, which we hypothesize requires reflection on the likely consequences of what one is about to say. The NFC provides a possible implementation for such reflection.

Importantly, the NFC's main function is not to produce creative behaviors that would surprise and delight a listener, but rather to avoid otherwise inevitable overextensions, and guard against productions that are likely to have unintended consequences. In other words, the NFC improves the precision of message transmission. That creative productions can then arise is a pleasant side effect.

6 Conclusion

This paper has proposed a production-internal definition of creativity – creativity involves the speaker suppressing the most accessible expression of a message in the moment of production (the first expression that comes to mind) and producing a less accessible expression that still expresses the intended message. This paper detailed how this process can take place within a connectionist framework for language production, and how it can give rise to innovative expressions like backformations and circumlocutions.

Importantly, the first expression that comes to mind tends to be the usual way of expressing the speaker's message in the speech community. Therefore, avoiding it usually leads to a production that is relatively novel and therefore likely to be perceived as creative by the listener. In this framework, all intentionally creative behavior such as writing a poem involves suppressing prepotent, habitual responses to a combination of message and context. However, much linguistic behavior that is not produced with the intention to be creative also involves the same mechanism of suppressing the first production that came to mind, and is therefore creative from a mechanistic, speaker-internal perspective. These behaviors (like backformations) are sometimes perceived to be creative, but the mechanism that produces them is the same whether they are perceived to be creative or not. In contrast, some behaviors that are often perceived to be creative are not creative from the speaker-internal perspective, because they do not involve suppression of a prepotent production.

For example, a child overextending *kitty* to mean COW is not being creative if they don't know the word *cow* and therefore have no alternative expression for the intended meaning (*cow*) to suppress. In contrast, a child who knows the word *cow* and accesses it first when naming the picture of a cow but opts to say *kitty* instead *is* being creative. Mechanistically-speaking, the first child simply allows activation from the message to spread to the associated forms and produces the first form that reaches a sufficient level of activation. The second child does something additional, suppressing the most active form (*cow*) and allowing the activation from the message to select another competitor. Creativity is about reaching the same intended destination by taking a path less traveled.

References

Albright, Adam. 2003. A quantitative study of Spanish paradigm gaps. *West Coast Conference on Formal Linguistics 22 Proceedings*. 1–14. Somerville, MA: Cascadilla Press.

Allopenna, Paul D., James S. Magnuson & Michael K. Tanenhaus. 1998. Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language* 38. 419–439.

Aronoff, Mark. 1976. *Word formation in generative grammar*. Cambridge, MA: MIT Press.

Berg, Thomas. 1998. *Linguistic structure and change: An explanation from language processing*. Oxford: Oxford University Press.

Boyd, Jeremy K., & Adele E. Goldberg. 2011. Learning what not to say: The role of statistical preemption and categorization in a-adjective production. *Language* 87. 55–83.

Brochhagen, Thomas, Gemma Boleda, Eleanora Gualdoni & Yang Xu. 2023. From language development to language evolution: A unified view of human lexical creativity. *Science* 381. 431–436.

Burridge, Kate. 2012. Euphemism and language change: The sixth and seventh ages. *Lexis. Journal in English Lexicology*. 7.

Buz, Esteban, Michael K. Tanenhaus, & T. Florian Jaeger. 2016. Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations. *Journal of Memory and Language* 89. 68–86.

Bybee, Joan L. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, Joan L. & Mary A. Brewer. 1980. Explanation in morphophonemics: changes in Provençal and Spanish preterite forms. *Lingua* 52. 201–242.

Bybee, Joan, & James L. McClelland. 2005. Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The Linguistic Review* 22. 381–410.

Bybee, Joan, & Dan Slobin. 1982. Why small children cannot change language on their own: Evidence from the English past tense. In A. Alqvist (ed.), *Papers from the 5th International Conference on Historical Linguistics*, 29–37. Amsterdam: John Benjamins.

Ceuppens, Hilke & Hendrik DeSmet. 2024. When does semantic change lead to semantic loss? Metaphor vs. inference-driven metonymy. Paper presented at the 57th Annual Meeting of the Societas Linguistica Europaea, Helsinki, Finland, August 21–24.

Chuang, Yu-Ying, Marie Lenka Vollmer, Elnaz Shafaei-Bajestan, Susanne Gahl, Peter Hendrix & R. Harlald Baayen. 2021. The processing of pseudoword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning. *Behavior Research Methods*, 53. 945–976.

Daland, Robert, Andrea D. Sims & Janet Pierrehumbert. 2007. Much ado about nothing: A social network model of Russian paradigmatic gaps. In *Proceedings of the 45th annual meeting of the Association of Computational Linguistics*. 936–943.

Davies, Mark. 2002. Un corpus anotado de 100.000.000 de palabras del Español histórico y moderno. *Procesamiento del Lenguaje Natural* 29.

Davies, Mark. 2012. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora* 7. 121–157.

Dell, Gary S. 1986. A spreading-activation theory of retrieval in sentence production. *Psychological Review* 93. 283–321.

Dell, Gary S. 1985. Positive feedback in hierarchical connectionist models: Applications to language production. *Cognitive Science* 9. 3–23.

Dhooge, Elisah & Robert J. Hartsuiker. 2011. How do speakers resist distraction? Evidence from a taboo picture-word interference task. *Psychological Science* 22. 855–859.

Diesburg, Darcy A., & Jan R. Wessel. 2021. The Pause-then-Cancel model of human action-stopping: Theoretical considerations and empirical evidence. *Neuroscience & Biobehavioral Reviews* 129. 17–34.

Fasmer, M. 1986. *Etymological dictionary of the Russian language*. Moscow: Progress.

Ferreira, Victor S. & Zenzi M. Griffin. 2003. Phonological influences on lexical (mis)selection. *Psychological Science* 14. 86–90.

Gershkoff-Stowe, Lisa & Linda B. Smith. 1997. A curvilinear trend in naming errors as a function of early vocabulary growth. *Cognitive Psychology* 34. 37–71.

Goldberg, Adele E. 1995. *Constructions*. Chicago: University of Chicago Press.

Goldberg, Adele E. & Fernanda Ferreira. 2022. Good-enough language production. *Trends in Cognitive Sciences* 26. 300–311.

Gorman, Kyle & Charles Yang. 2019. When nobody wins. In Franz Rainer, Francesco Gardani, Wolfgang U. Dressler, & Hans Christian Luschützky (eds.), *Competition in inflection and word-formation*, 169–193. Cham: Springer.

Gries, Stefan T. & Nick C. Ellis. 2015. Statistical measures for usage-based linguistics. *Language Learning* 65. 228–255.

Grishina, Elena. 2006. Spoken Russian in the Russian National Corpus (RNC). In *Proceedings of LREC*, 121–124. Genoa: International Conference on Language Resources and Evaluation.

Halle, Morris. 1973. Prolegomena to a theory of word formation. *Linguistic Inquiry* 4. 3–16.

Harmon, Zara & Vsevolod Kapatsinski. 2017. Putting old tools to novel uses: The role of form accessibility in semantic extension. *Cognitive Psychology* 98. 22–44.

Harmon, Zara & Vsevolod Kapatsinski. 2021. A theory of repetition and retrieval in language production. *Psychological Review* 128. 1112–1144.

Hartsuiker, Robert J. & Herman H. J. Kolk. 2001. Error monitoring in speech production: A computational test of the perceptual loop theory. *Cognitive Psychology* 42. 113–157.

Heinen, David JP, & Dan R. Johnson. 2018. Semantic distance: An automated measure of creativity that is novel and appropriate. *Psychology of Aesthetics, Creativity, and the* Arts 12. 144.

Hinton, Geoffrey, James L. McClelland, & David E. Rumelhart. 1986. Distributed representations. In David E. Rumelhart, James L. McClelland, & the PDP Research Group (eds.), *Parallel distributed*

processing. Explorations in the microstructure of cognition. Vol. 1: Foundations, 77–109. Cambridge, MA: MIT Press.

Hoeffner, James H. & James L. McClelland. 1993. Can a perceptual processing deficit explain the impairment of inflectional morphology in developmental dysphasia? A computational investigation. In *Proceedings of the 25th Annual Child Language Research Forum*. Stanford, CA: CSLI.

Hoffmann, Thomas. 2019. Language and creativity: A Construction Grammar approach to linguistic creativity. *Linguistics Vanguard* 5. 20190019.

Holler, Judith, Phillip M. Alday, Caitlin Decuyper, Mareike Geiger, Kobin H. Kendrick & Antje S. Meyer. 2021. Competition reduces response times in multiparty conversation. *Frontiers in Psychology* 12. 693124.

Kapatsinski, Vsevolod. 2009. Adversative conjunction choice in Russian (no, da, odnako): Semantic and syntactic influences on lexical selection. *Language Variation and Change* 21. 157–173.

Kapatsinski, Vsevolod. 2017. Copying, the source of creativity. In Anna Makarova, Stephen Dickey & Dagmar Divjak (eds.), *Each venture a new beginning: Studies in honor of Laura A. Janda*, 57–70. Bloomington, IN: Slavica.

Kapatsinski, Vsevolod. 2018. *Changing minds changing tools: From learning theory to language acquisition to language change*. Cambridge, MA: MIT Press.

Kapatsinski, Vsevolod. 2021. What are constructions, and what else is out there? An associationist perspective. *Frontiers in Communication* 5. 575242.

Kapatsinski, Vsevolod. 2022. Morphology in a parallel, distributed, interactive architecture of language production. *Frontiers in Artificial Intelligence* 5. 803259.

Kapatsinski, Vsevolod & Zara Harmon. 2017. A Hebbian account of entrenchment and (over)-extension in language learning. In Glenn Gunzelmann, Andrew Howes, Thora Tenbrink & Eddy Davelaar (eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, 2366–2371. Austin: Cognitive Science Society.

Koranda, Mark J., Martin Zettersten & Maryellen C. MacDonald. 2022. Good-enough production: Selecting easier words instead of more accurate ones. *Psychological Science* 33. 1440–1451.

Langacker, Ronald W. 1987. *Foundations of cognitive grammar: Volume I: Theoretical prerequisites*. Stanford, CA: Stanford University Press.

Martin, Andrew T. 2007. *The evolving lexicon*. Ph.D. Dissertation, UCLA.

McClelland, James L. 1981. Retrieving general and specific information from stored knowledge of specifics. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *3*, 170–172.

Motley, Michael T., Carl T. Camden & Bernard J. Baars. 1982. Covert formulation and editing of anomalies in speech production: Evidence from experimentally elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior* 21. 578–594.

Naigles, Letitia G. & Susan A. Gelman. 1995. Overextensions in comprehension and production revisited: Preferential-looking in a study of *dog*, *cat*, and *cow*. *Journal of Child Language* 22. 19–46.

Nozari, Nazbanou, Gary S. Dell & Myrna F. Schwartz. 2011. Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production. *Cognitive Psychology* 63. 1–33.

Nozari, Nazbanou & Christopher R. Hepner. 2019. To select or to wait? The importance of criterion setting in debates of competitive lexical selection. *Cognitive Neuropsychology* 36. 193–207.

Oldfield, Richard & Arthur Wingfield. 1965. Response latencies in naming objects. *Quarterly Journal of Experimental Psychology* 17. 273–281.

Pirog Revill, Kathleen, Richard N. Aslin, Michael K. Tanenhaus & Daphne Bavelier. 2008. Neural correlates of partial lexical activation. *Proceedings of the National Academy of Sciences* 105. 13111–13115.

Ramscar, Michael, Melody Dye & Joseph Klein. 2013. Children value informativity over logic in word learning. *Psychological Science* 24. 1017–1023.

Rastelli, Clara, Antonino Greco, Nicola De Pisapia & Chiara Finocchiaro. 2022. Balancing novelty and appropriateness leads to creative associations in children. *PNAS Nexus* 1. pgac273.

Rogers, Timothy T. & James L. McClelland. 2004. *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Sampson, Geoffrey. 2016. Two ideas of creativity. In Martin Hinton (ed.), *Evidence, experiment and argument in linguistics and philosophy of language*, 15–26. Bern: Peter Lang.

Savčuk, S. O., Arxangelskij, T. A., Bonč-Osmolovskaja, A. A., Donina, O. V., Kuznecova, Yu. N., Ljashevskaja, O. N., Orexov, B. V. & Podrjadčikova, M. V. 2024. Nacional'nyj korpus russkogo jazyka 2.0: Novye vozmožnosti i perspektivy razvitija. [National corpus of Russian 2.0: New capacities and prospects for development.] *Voprosy Jazykoznanija* 2. 7–34.

Schertz, Jessamyn. 2013. Exaggeration of featural contrasts in clarifications of misheard speech in English. *Journal of Phonetics* 41. 249–263.

Schwartz, Richard G. & Lawrence B. Leonard. 1982. Do children pick and choose? An examination of phonological selection and avoidance in early lexical acquisition. *Journal of Child Language* 9. 319–336

Sims, Andrea D. 2015. *Inflectional defectiveness*. Cambridge: Cambridge University Press.

Stankiewicz, Edward. 1957. The expression of affection in Russian proper names. *The Slavic and East European Journal* 1. 196–210.

Staub, Adrian. 2009. On the interpretation of the number attraction effect: Response time evidence. *Journal of Memory and Language*, 60 (2). 308–327.

Stern, Michael C. & Jason A. Shaw. 2023. Neural inhibition during speech planning contributes to contrastive hyperarticulation. *Journal of Memory and Language* 132. 104443.

Suttle, Laura & Adele E. Goldberg. 2011. The partial productivity of constructions as induction. *Linguistics* 49. 1237–1269.

Teruya, Hideko & Vsevolod Kapatsinski. 2019. Deciding to look: Revisiting the linking hypothesis for spoken word recognition in the visual world. *Language, Cognition and Neuroscience* 34. 861–880.

Tiersma, Peter Meijes. 1982. Local and general markedness. *Language* 58. 832–849.

Trask, Larry. 1996. *Historical linguistics*. London: Arnold.

Traugott, Elizabeth Closs. 1988. Pragmatic strengthening and grammaticalization. In *Annual Meeting of the Berkeley Linguistics Society*. 406–416.

Wedel, Andrew, Noah Nelson & Rebecca Sharp. 2018. The phonetic specificity of contrastive hyperarticulation in natural speech. *Journal of Memory and Language* 100. 61–88.

Wood, Wendy, and David T. Neal. 2007. A new look at habits and the habit-goal interface." *Psychological Review* 114. 843–861.

Yee, Eiling & Julie C. Sedivy. 2006. Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 32. 1–14.