Squib

On Natural Language Stringsets, Senses, and Generative Capacity

Daniel Radzinski

radzinski@gmail.com

It is demonstrated that the proofs given in prominent and well-established weak generative capacity arguments for natural language are flawed, due to unexpected interpretations of strings. However, once unique representations of lexical semantic senses form part of such intersection-based proofs, the arguments stand.

1. Introduction

The results presented by Huybregts (1984) and Shieber (1985)[1] regarding cross-serial verb raising (**VR**) exhibited in Swiss German (**SwG**) have become the typically cited sources that "prove" the non-context freeness of natural language (**NL**). Manaster Ramer (1988, page 101) and Ojeda (1988, pages 488–489) clarified a crucial formal failure in Shieber's argument—while also providing a correction to it—yet these reservations have gone largely unnoticed and been effectively ignored. The supposed correctness of Huybregts and Shieber's (**H&S**) findings has been taken mostly for granted over the years. We find textbooks like Partee et al. (1990, page 503), for example, stating that the evidence from Shieber (1985) "seems unassailable on either formal or empirical grounds." Likewise, Jurafsky and Martin (2000, page 489) describe the H&S argument as "correct (or at least not-yet disproved)" and as a "successful proof". More recently, Jäger and Rogers (2012, page 1960) indicate with respect to "[t]he issue of whether all natural languages are context-free that [i]t was finally settled only in the mid-1980s, independently by the scholars Riny Huybregts [...], Stuart Shieber [...] and Christopher Culy."[2] And, even more recently, Dolatian et al. (2021, page 229) categorize SwG as being context free (**CF**) in terms of strong generative capacity (**SGC**) as well as weak generative capacity (**WGC**)—whereas Dutch is non-CF only in SGC—on the basis of Shieber (1985), and Wang and Steinert-Threlkeld (2023, page 272) point out that whether CF grammars "could serve as a computational model for natural language, ... had been

[1] Huybregts and Shieber published their findings independently at roughly the same time. The linguistic and formal argumentation regarding the Swiss German facts is effectively the same in both and will henceforth be treated as a single argument. Unless stated otherwise in this squib, focus will be on Shieber's version, as it is the wider-known one.

2 Culy (1985) gave a non-context freeness proof for the vocabulary of Bambara, a West African Mande language.

Action Editor: Michael White. Submission received: 13 March 2025; revised version received: 22 May 2025; accepted for publication: 30 June 2025.

https://doi.org/10.1162/COLI.a.21

© 2025 Association for Computational Linguistics
Published under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license

an open question for a few decades until it was settled [in 1985 -DR] by evidence such as Swiss German cross-serial dependency . . . ".

There is, however, a serious flaw in the H&S argument that invalidates it, and which until recently appears never to have been raised specifically in connection with it. In fact, the flaw renders many, if not all or most, WGC arguments that have been raised against CFness and against membership in other mildly context sensitive classes also invalid. Section 2 elaborates on that flaw. Section 3 proposes a different approach, invoking lexical semantic considerations when giving WGC proofs of NL, including English, SwG, Dutch, and Chinese. Section 4 presents the conclusions of this study.

2. Unexpected Interpretations

H&S presented VR phenomena in SwG rather than Dutch because the former exhibits overt morphological case while the latter does not. This limits the argument to a domain purely within syntax with no need to deal with consequences of selectional restrictions, allowing in principle for the formulation of a valid non-CFness proof, paraphrased as:

> Let the regular language (**RL**) R be {*Jan säit das mer d'chind* (em Hans)* es huus haend wele laa* hälfe* aastriche*}. The intersection of R with SwG, or, more precisely, with a formalized idealization of the NL SwG, is T = {*Jan säit das mer d'chind*^{m} *(em Hans)*^{n} *es huus haend wele laa*^{j} *hälfe*^{k} *aastriche*: $1 \leq m \leq j$ and $1 \leq n \leq k$}.[3] T is provably non-CF by contradiction via application of the strong pumping lemma for CF languages (**CFLs**). CFLs are closed under intersection with RLs. R is regular and T is non-CF, therefore SwG is also non-CF.

This proof is based on certain acceptability judgments within SwG. For example, sentence (1) below is acceptable, while sentence (2) should not be:[4]

(1) *Jan säit das mer d'chind em Hans es huus haend wele laa hälfe aastriche*
Jan says that we ACC- the children DAT-Hans ACC-the house have wanted let help paint
Jan says that we wanted to let the children help Hans paint the house.

(2) *Jan säit das mer d'chind em Hans em Hans es huus haend wele laa hälfe aastriche*
Jan says that we ACC-the children DAT-Hans DAT-Hans ACC-the house have wanted let help paint

The unacceptability of sentence (2) is because the total number of instances of the dative noun phrase *em Hans* cannot exceed the total number of instances of the dative-subcategorizing verb *hälfe*, as we see explicitly in T, which is the intersection of the RL R with SwG. So, sentence (2) seemingly makes no sense. But what if *Hans em Hans* were the name of an individual? Such a name might sound silly, unconventional, and unexpected perhaps, yet it would nonetheless be a possible, legitimate proper name. Then sentence (2) would make perfect sense as "Jan says that we wanted to let the children help Hans em Hans paint the house." Based on this acceptance alone, the intersection of R with SwG is not the non-CFL T, but rather U = {*Jan säit das mer d'chind*^{m}

3 T is according to Ojeda's (1988, page 488) aforementioned correction to Shieber's proof (due to Alexis Manaster Ramer). The correction reflects the linguistic fact that SwG allows for subjects of infinitives in VR cases to be optionally omitted. Per Shieber's erroneous intersection, T would have been {*Jan säit das mer d'chind*^{m} *(em Hans)*^{n} *es huus haend wele laa*^{m} *hälfe*^{n} *aastriche: m, n* ≤ 1}, also a non-CFL.

4 ACC stands for accusative case marker; DAT stands for dative case marker.

(em Hans)* *es huus haend wele laa*^{j} *hälfe** *aastriche*: $1 \leq m \leq j\}$, which is CF.[5] Consequently, the allegedly unassailable H&S proof is vitiated. Stabler (2019, page 249) stated: "Shieber (1985) claims that the string set of Swiss German is not in CFLs, but he pays no attention to how this claim is threatened by direct quotation." Not only is this claim threatened specifically by direct quotation, but, as we see in the aforementioned proof, by proper names in general.

The notion that a seemingly artificial, innocuous proper name like *Hans em Hans* could serve as the "culprit" that disproves a decades-long accepted argument might make many feel uneasy at first, to say the very least. Nevertheless, unfortunately perhaps, we must accept certain strings, such as proper names, as legitimate, due to the many unexpected meanings they may have. *Hans em Hans* is one such case. As aptly indicated first by Pullum and Gazdar (1982, page 490): "so many strings turn out to be grammatical by accident under an unintended interpretation" and later by Manaster Ramer (1987, page 234): "any string in any language is likely to have a multitude of syntactic structures (and meanings), most of them highly artificial and unlikely to occur under natural conditions, but theoretically legitimate." And, more recently, Stabler (2019, page 244) wrote regarding the oddness, but perfect intelligibility and grammaticality, of the quotation name *Mat the on is cat the* that "[e]very human language allows direct quotation, so in every language, every sequence of words has a grammatical structure."[6]

It is precisely due to unexpected cases of possible ambiguity that Manaster Ramer's (1987) argument against the CFness of Dutch (and standard German), on the basis of VR *cum* constituent coordination, ended up being reformulated not in terms of WGC, but rather classificatory capacity (**CC**). Manaster Ramer (1987, page 238) preliminarily defined CC "as the measure of a formalism's ability [to] classify a set of strings (and substrings) and specify which ones are like which other ones." Likewise, Radzinski (1990, 1991) invoked CC considerations, the former in an argument against the CFness of Chinese based on A-not-A questions and the latter, in an argument that Chinese was not a Multiple CFL (**MCFL**) based on its arbitrarily long strung-together number-names.[7] As intuitively clear as CC may be, Manaster Ramer's definition of this notion was only partially refined later in Radzinski (1990, page 122), but never elaborated any further or fully defined formally.

5 U is homomorphic to $\{a^n b^m : 1 \leq n \leq m\}$, a well-known CFL.

6 While the content of Stabler's quotation name might seem odd and farfetched when taken as a proper name, it suffices to peruse a list of actual horse names, e.g., `https://www.bloodhorse.com/horse-racing/thoroughbred-racing/leaders/horses/2017`, to see, inter alia, a wide variety of oddly formed names: *Practical Joke* (NP); *Gun Runner* (compound noun); *Mind Your Biscuits* (VP, full sentence); *It Tiz Well* (made-up alternative spelling for "It is well", full sentence); *Hence* (adverb); *By the Moon* (PP); *La Coronel* (NP in foreign language with gender disagreement); *Bolt d'Oro* (language mix); *Paulassilverlining* (possessive sans apostrophe and spaces); *Itsinthepost* (full sentence sans apostrophe and spaces); *Matt King Coal* (facetious take on well-known musician), and many more. Such free variation in acceptable names suggests that the set of possible proper names may in fact comprise Σ^*, i.e., all string combinations within a language, making the WGC of NL regular and effectively uninteresting. However, as will be shown in Section 3 below, even if the set of *bare* NL strings is Σ^*, this does not quite render NL regular and its WGC trivial or pointless.

7 Kanazawa (2009) addressed a formal failure in the conclusion of the pumping lemma used by Radzinski (1991) to show that Chinese number-names formed a non-MCFL. That lemma, an erroneously concluded variant of Seki et al.'s (1991, pages 200–201) pumping lemma, might perhaps be strong enough to show that the number-names are outside of the class of Well-Nested MCFLs, but not necessarily outside of the larger MCFL class. At the same time, however, Kanazawa indicated that Michaelis and Kracht (1997) had shown that Chinese number-names were non-semilinear, hence, a fortiori a non-MCFL. So, the claim raised in Radzinski (1991) regarding the exclusion of Chinese number-names from the class of MCFLs remains proven.

Most other WGC arguments prior to, during, and after H&S have effectively ignored unexpected interpretations, despite the difficulties they create as shown above, in terms of formulating a sound and rigorous mathematical proof backed by accurate linguistic facts. In other words, the possible issues stemming from ambiguities may have been known to and understood by the proponents of WGC arguments, yet they all along focused only on one meaning of the strings in question, allowing their arguments to go through. The next section expands on why such a focus may be warranted, after all, and suggests a means for making it explicit and applying it aptly within WGC arguments.

3. Senses

Hopcroft and Ullman (1979, pages 1–2) define a **string** as "a finite sequence of symbols juxtaposed" and an **alphabet** as "a finite set of symbols." Such definitions may work rather well with formal languages, but less so with NLs. This is because the alphabet (or vocabulary) of an NL does not comprise merely symbols of concatenated sounds, or an orthographic representation of such, but rather symbols consisting of bare words and their concomitant meaning, or *sense*, something linguists have understood for quite some time. Miller (1995, page 39) expresses this rather clearly with respect to NLs:

> We define the vocabulary of a language as a set W of pairs (f,s), where a form f is a string over a finite alphabet, and a sense s is an element from a given set of meanings. Forms can be utterances composed of a string of phonemes or inscriptions composed of a string of characters. Each form with a sense in a language is called a *word* in that language.

Let's then follow Miller's definition and represent NL strings accordingly. This can be achieved, for example, by concatenating the word's spelling with a formal representation of its sense. The sense can be represented by a unique identifier associated with the word's lexical formal concept, as organized, for example, in Miller's lexical knowledge base WordNet (`http://wordnetweb.princeton.edu/perl/webwn`) or other authoritative concept-based controlled vocabulary (dictionary, lexicon, thesaurus, etc.) that includes a representation of semantic relations.[8] WordNet assigns fixed unique IDs to the senses of a word in the form of a Database Location (**DL**) integer. This sense key is semantic in nature and differs from a word's part-of-speech (**POS**) category, which is syntactic in nature. Together with a word's orthography and semantic field, both form a unique relational database key for every word in WordNet. Concatenating a bare word with one of its DL IDs explicitly yields a word coupled with its intended sense, excluding any other possible interpretation of that word. Accordingly, the alphabet, or vocabulary, of English, for example, would consist of symbols such as bag02776042 (glossed as 'a flexible container with a single opening'), box02886585 (glossed as 'a (usually rectangular) container; may have a lid'), etc., rather than just plain "bag" or "box". In other words, the sense of a word gets encoded explicitly in the graphical representation of that word by means of a unique sense identifier appended to the standard orthography of the word.

8 Varying granularities found across such concept-based controlled vocabularies when differentiating between senses do not affect our claims (but see footnote 16 below).

Now, let's take a look at a WGC argument for English from the present century. Pullum and Rawlins (2007) examined, with arguable conclusions, a non-CFness argument for English, based on adjuncts of the form *X or no X*, e.g., "The North Koreans were developing nuclear weapons anyway, Iraq war or no Iraq war". Their formal argumentation rests on intersecting the RL R, where $R = \{$We will do it, X_1 or no $X_2 | X_1, X_2 \in \{\text{box, bag}\}^+\}$,[9] with English. The result of their intersection is {We will do it, X or no $X \mid X \in \{\text{box, bag}\}^+\}$, which is non-CF, therefore, neither is English. This is under their assumption that sentence (3) below is acceptable, while sentence (4) is not:

(3) The show will go on, box bag bag or no box bag bag.

(4) The show will go on, box bag bag or no box bag box.

Nevertheless, sentence (4) is actually perfectly fine, under the assumption that "box bag bag or no box bag box" is, for example, a nickname for Mister Ed and his owner Wilbur is telling that horse that they are going to perform some action. Again, unexpected interpretations rear their ugly head. So, the result of the intersection of R with English is R itself, an RL. In fact, once we take strings in NL to be composed of their orthographic form concatenated with their sense ID, as presented above, then the result of the intersection is really the null set, since "box" and "bag" are not in English, in contrast to box02886585, box02887466, bag13776918, etc. A more insightful intersection would be between English and the RL RE where RE = {we11111 will22222 do01716563 it33333, X_1 or44444 no02276242 $X_2 \mid X_1, X_2 \in \{\text{box02886585, bag02776042}\}^+$}.[10] This intersection yields {we11111 will22222 do01716563 it33333, X or44444 no02276242 $X \mid X \in \{\text{box02886585, bag02776042}\}^+$}, the intended and desired non-CF result (regardless of any other conclusions reached later in that paper). This is because box02886585 is a string representing "box" *only* in the sense of 'a (usually rectangular) container; may have a lid' and bag02776042, one representing "bag" *only* in the sense of 'a flexible container with a single opening'. Having these two be part of the RL used in the intersection with English excludes any intervention from other senses of the bare words "box" and "bag", such as box02887466 in the sense of 'private area in a theater or grandstand where a small group can watch the performance' and bag13776918 in the sense of 'the quantity of game taken in a particular period (usually by one person)'. It also excludes the (proper) name sense these bare words exhibit when they form part of a nickname such as "box bag bag or no box bag box". Such a case would be represented by a string along the lines of "box01010 bag01010 bag01010 or01010 no01010 box01010 bag01010 box01010", where 01010 would be an ID convention for any bare word used solely in a name-formation context. This 01010 convention applies to any word and needn't be listed at all in WordNet, or any other controlled vocabulary. The idea behind this is that words that serve as a token of a name are devoid of any of their possible *listed* senses. Their sense is effectively something along the lines of "part of name" and

9 The beginning of their RL is actually "We'll". The non-contracted form "We will" is used here instead, for the sake of simplicity of exposition.

10 Some of these DL IDs are innocuously improvised, since WordNet focuses on nouns, verbs, adjectives, and adverbs, excluding closed class words, but clearly not all. The point here is that a bare word should be concatenated with a unique integer reflecting its intended sense. It is preferable, of course, that such integer be extracted from an authoritative system such as WordNet, but if absent therein, an improvised unique integer can do just as well.

this sense can apply to any word, listed in a controlled vocabulary or not.[11] Thus, the senses of the three instances of "box" in Mister Ed's potential nickname of "box bag bag or no box bag box" are neither WordNet's 02886585: 'a (usually rectangular) container; may have a lid' nor 02887466: 'private area in a theater or grandstand where a small group can watch the performance' (nor any of the other WordNet senses associated with "box"), but rather the proposed 01010, a dummy of sorts assigned as a sense for each word of a name. Ditto for the instances of "bag", mutatis mutandis. This reformulated argument for *X or no X* effectively circumvents any of the unwanted effects caused by unexpected interpretations and leads to a non-CF result for English, assuming arguable appropriate support stemming from the relevant linguistic facts.

Once WGC arguments are formulated in a way that reflects the representation of NL strings as consisting of bare words coupled with their specific senses, many of them cease to be vitiated by the effects of unexpected interpretations, similarly to what we saw above in the case of Pullum and Rawlins (2007). This is true for a reformulated H&S SwG argument (since the *em* in the name *Hans em Hans* has a different unique sense ID than that of the dative marker *em*), so SwG is indeed weakly non-CF. But then, pace Dolatian et al. (2021, page 229), *so is Dutch*! The overt morphological case markings that SwG exhibits, but Dutch does not, no longer play any role in determining whether an NL is weakly CF or, perhaps, only strongly so. It is purely SwG words composed of orthography-*cum*-sense that determine this in a formal WGC proof. The same holds for Dutch. Pullum and Gazdar's (1982, page 488) claim to the effect that "*(dat Jan) Marie Arabisch Pieter wil laten zien schrijven* '(that Jan) will let Marie see Arabic write Pieter' is perfect under the assumption that there is a language or writing system called 'Pieter' and a person named 'Arabisch' has learned to write it" can no longer be used to weaken a non-CF argument for Dutch, since *Arabisch* as the name of a person differs from *Arabisch* with the sense of *taal v.d. Arabieren* 'language of the Arabs'. Manaster Ramer's (1987, pages 228–233) argument for the non-CFness of Dutch, for example, rests on acceptability judgments of sentences like (5) and (6) below (where *Of* serves as an echo question marker):

(5) *Of Maria Johannes Mathijs zag leren zwemmen?*
Maria Johannes Mathijs saw teach swim
Did Maria see Johannes teach Mathijs to swim?

(6) *Of Maria Johannes Mathijs zag zwemmen?*
Maria Johannes Mathijs saw swim

Sentence (5) is perfectly acceptable while sentence (6) is seemingly not, due to its lack of correlation between the number of initial proper names and subsequent

11 This approach has been described by an anonymous reviewer as one in which the senses of name parts "are novel uses divorced from their old senses", a perfectly legitimate approach. At the same time, however, we must keep in mind that there exist alternative approaches, typically compositional, e.g., Pagin and Westerståhl (2010), in which the original senses are maintained. Adopting these latter approaches might lead to results that differ from those in this squib in terms of obtaining successful reformulations of WGC arguments.

serialized verbs. This is true under the assumption that the relevant senses in sentence (6) are:

Maria480510392 Johannes995428834 Mathijs074345384 zag10253 zwemmen10335[12]

However, sentence (6) is perfectly fine once Johannes Mathijs serves as the name of a single individual, i.e., once the senses in the sentence are:

Maria480510392 Johannes01010 Mathijs01010 zag10253 zwemmen10335

Here, Johannes01010 and Mathijs01010 are tokens of a single proper name, Johannes Mathijs, thus maintaining an appropriate correlation between the number of proper names and the number of serialized verbs required for Dutch VR serialization. Manaster Ramer (1987, page 236) vitiated his earlier claim, due to the acceptable interpretation we notice sentences like (6) can have. However, once his argument is reformulated with an intersection-based proof using orthography-*cum*-sense words, instead of bare words, he needn't be concerned at all with such acceptability, as the unintended senses involved no longer intervene in the proof. Therefore, his argument does indeed hold on WGC grounds with no need to resort to CC considerations.[13] Likewise, Radzinski (1991, page 293) raised a concern that *wu zhao wu zhao zhao* 'five trillion five trillion trillion' could be a "well-formed proper name, such as the title of a book, for example", (in contrast to *wu zhao zhao wu zhao* 'five trillion trillion five trillion', a well-formed number-name) even though it may not be a well-formed Chinese number-name. As such, the WGC non-MCFL proof raised earlier in that article would hold only for Numeric Chinese and not for Chinese in its entirety. Chinese would be a non-MCFL only by CC considerations. Nevertheless, once senses are considered, wu13563536 with the sense of 'the cardinal number that is the sum of four and one', used, inter alia, for number-name formation purposes is not the same as wu01010 used for general name-formation purposes. Ditto re zhao13571318 with the sense of 'the number that is represented as a one followed by 12 zeros' vs. zhao01010 used only for general name-formation purposes. So, the intersection of the RL $\{(\text{wu13563536 zhao13571318}^{+})^{+}\}$ with Chinese—*not merely Numeric Chinese, but Chinese in its entirety*—is $\{\text{wu13563536 zhao13571318}^{k1}\ \text{wu13563536 zhao13571318}^{k2}\ \ldots\ \text{wu13563536 zhao13571318}^{kn} \mid \text{k1} > \text{k2} > \ldots > \text{kn} > 0\}$, a non-semilinear language per Michaelis and Kracht (1997), a fortiori a non-MCFL. So, Chinese, as a full NL, is not well-behaved semilinear, with no need to appeal to any CC considerations.[14]

Notice that nowhere in this proposal have I alluded to any trees, structures, or POS—therefore, by no means are we dealing here with SGC, but rather purely with

12 The sense ID numbers used here are the actual sense ID numbers for these words in Open Dutch WordNet (`https://raw.githubusercontent.com/cltl/OpenDutchWordnet/refs/heads/master/resources/cili/odwn_cili.xml`).

13 Resting on similar judgments, Manaster Ramer (1987, page 233) also argued against the membership of Dutch in the classes of Tree Adjoining Languages and Multi-Component Tree Adjoining Languages (equivalent to MCFLs). Groenink (1997, pages 612–613) gave the formal proofs to such claims. On the basis of Groenink's proofs, Michaelis and Kracht (1997) extend the argument to show that Dutch is not semilinear (provided that certain properties of coordination phenomena in that language hold).

14 This result for Chinese also has implications for the status of English as a semilinear language, the discussion of which is beyond the scope of this short squib. Per Ibarra and McQuillan (2020), semilinear languages closed under homomorphism, inverse homomorphism, and intersection with RLs are said to be "well-behaved."

WGC.[15] In fact, the definitions for the many grammars that generate tree structures remain largely unchanged. For example, Hopcroft and Ullman's (1979, page 79) denotation of a CF grammar as consisting of a quadruple (V, T, P, S), where V is a finite set of non-terminals, T is a finite set of terminals, P is a finite set of production rules, and S is a single start symbol from V, remains the same. The terminals, or leaves, take the form of words like bag02776042, rather than "bag", but otherwise nothing would change in a CF grammar for English (regardless for a moment of the status of English as a non-CFL). The terminals, of course, reflect a word's spelling in its first part and its unique sense, in its second part. This is how linguists have typically perceived the notion of a word. As long as it is only terminals that are referenced in WGC arguments, such arguments remain in the realm of WGC and not SGC. SGC arguments rely crucially on phrasal and sentential trees as assigned by grammar theories. In such cases, the trees under discussion are based on the predictions of a subjectively chosen framework. Other frameworks might posit very different structures for the same sentence(s) under discussion, turning the arguments into ones highly dependent on whether one accepts the assigned tree structures for the relevant sentences or not. WGC, on the other hand, remains agnostic to any specific grammatical theory. It is, therefore, preferable to formulate generative capacity arguments as ones of WGC, rather than SGC.[16]

Notwithstanding, one cannot ignore the fact that making implicit or explicit reference to word senses, which in contrast to plain orthographic forms are not observable in a text, adds a layer of information typically unavailable to NL recognition/parsing tasks (except when successful WSD takes place before any syntactic recognition/parsing does). As such, some might view a proposal of formulating WGC arguments using orthography-*cum*-sense as departing from traditional practice. Yet, as shown above, due to the nature of names and unexpected interpretations, focusing squarely on orthography-*sans*-sense requires no more than finite-state machinery for recognition or parsing purposes. Sensible parsing, reflecting at a minimum some plausible semantic interpretation, requires more powerful mechanisms and this is precisely the conclusion typically reached by WGC arguments/proofs. Therefore, as indicated, by ignoring unexpected senses the proponents of such arguments have already been acting implicitly in the orthography-*cum*-sense world, which includes unobservable evidence. So, it remains unclear whether there actually is any departure from traditional practice in the current proposal, other than making the implicit explicit.

15 Anonymous reviewers have expressed concern that given the correlation word-sense disambiguation (**WSD**) may have with POS tagging, explicit sense marking within proofs might entail implicit reference to POS in the proof, effectively moving us into SGC territory. In other words, if WSD is performed inter alia by access to POS information, e.g., in Zhong and Ng (2010), then using a specific sense ID, thus ruling out other potentially ambiguous senses, implies taking advantage of syntactic POS, which many would argue is in the realm of SGC. Nevertheless, high quality WSD needn't necessarily access any POS information at all, as we may find in models taking advantage of embeddings instead, e.g., Melamud et al. (2016). So, using sense information on its own does not necessarily entail access to syntactic POS information. Moreover, if one insists, after all, on considering the mere use of senses as falling within SGC, then previous generative capacity arguments classified as WGC ones have in fact been actually SGC arguments as well, since, as pointed out, particular senses were always implied in the intersection-based proofs of these arguments in order to exclude unexpected interpretations from consideration. Making sense assumptions explicit in a generative capacity argument, as proposed in this squib, does not alter the argument's generative capacity status (between weak and strong).

16 Once senses form an inherent part of NL stringsets, as, for example, per Miller's aforementioned definition, we ought to keep in mind a possible subjectivity due to different granularities of the senses of a particular word, depending on which controlled vocabulary is chosen. However, such subjectivity has virtually no effect on intersection-based proofs that use bare strings coupled with their word sense, if such senses effectively exclude any unwanted strings from consideration (as expected from the use of senses stemming from any such controlled vocabulary).

4. Conclusions

To recap:

(a) Shieber's (1985) proof as given for the weak non-context freeness of Swiss German is flawed. The same holds for Huybregts (1984) and for other weak generative capacity proofs in other languages that ignore unexpected interpretations of strings.

(b) Once words in natural language are represented by their orthography combined with an explicit unique identifier for their particular sense, arguments can be reformulated differently and successfully, including Huybregts (1984) and Shieber (1985), avoiding the effects of unexpected interpretations. Proofs within such reformulated arguments are in the realm of weak generative capacity and not strong generative capacity, as they involve neither trees nor parts-of-speech.

(c) In light of (b), Dutch is just as weakly non-context free as Swiss German is, due to its cross-serial verb raising, regardless of the latter having overt case markers and the former, not. It is also not semilinear, per Michaelis and Kracht (1997).

(d) In light of (b), and following a reformulation of Radzinski (1991), Chinese, rather than merely Numeric Chinese, is not well-behaved semilinear, a fortiori not context free, tree adjoining, or multiple context free, due to its long strung-together number-names.

The innovation of integrating formally assigned senses within the representation of words used in weak generative capacity arguments makes explicit a long-held implicit understanding that linguists have had regarding natural language strings. While we can't expect all future generative capacity arguments to be formulated as done in this squib, disclaimers to the effect that natural language words or strings in an intersection-based proof should be understood as including sense content would be in order.

Acknowledgments

I am grateful to Marcus Kracht, Geoff Pullum, Stuart Shieber, and Mike White, as well as anonymous reviewers, for their remarks on earlier drafts of this squib. I'd also like to take the opportunity to thank Polly Jacobson for bringing to my attention three and a half decades ago the infelicity of ignoring semantic considerations when focusing on issues related to the generative capacity of natural language.

References

Culy, Christopher. 1985. The complexity of the vocabulary of Bambara. *Linguistics and Philosophy*, 8:345–351. https://doi.org/10.1007/BF00630918

Dolatian, Hossep, Jonathan Rawski, and Jeffrey Heinz. 2021. Strong generative capacity of morphological processes. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 228–243. https://doi.org/10.7275/sckf-8f46

Groenink, Annius V. 1997. Mild context-sensitivity and tuple-based generalizations of context-free grammar. *Linguistics and Philosophy*, 20:607–636. https://doi.org/10.1023/A:1005376413354

Hopcroft, John E. and Jeffrey D. Ullman. 1979. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley.

Huybregts, Riny. 1984. The weak inadequacy of context-free phrase structure grammars.

In Ger de Haan, Mieke Trommelen, and Wim Zonneveld, editors, *Van Periferie naar Kern*, pages 81–99. Foris.

Ibarra, Oscar H. and Ian McQuillan. 2020. Semilinearity of families of languages. *International Journal of Foundations of Computer Science*, 31:1179–1198. https://doi.org/10.1142/S0129054120420095

Jäger, Gerhard and James Rogers. 2012. Formal language theory: refining the Chomsky hierarchy. *Philosophical Transactions of the Royal Society B*, 367:1956–1970. https://doi.org/10.1098/rstb.2012.0077, PubMed: 22688632

Jurafsky, Daniel and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall.

Kanazawa, Makoto. 2009. The pumping lemma for well-nested multiple context-free languages. In V. Diekert and D. Nowotka, editors, *Developments in Language Theory: 13th International Conference, Lecture Notes in Computer Science*, volume 5583, pages 312–325. Springer. https://doi.org/10.1007/978-3-642-02737-6_25

Manaster Ramer, Alexis. 1987. Dutch as a formal language. *Linguistics and Philosophy*, 10:221–246. https://doi.org/10.1007/BF00584319

Manaster Ramer, Alexis. 1988. Review of Walter J. Savitch, Emmon Bach, William Marsh, and Gila Safran-Naveh, editors, *The Formal Complexity of Natural Language. Computational Linguistics*, 14:98–103.

Melamud, Oren, Jacob Goldberger, and Ido Dagan. 2016. Context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61. https://doi.org/10.18653/v1/K16-1006

Michaelis, Jens and Marcus Kracht. 1997. Semilinearity as a syntactic invariant. In Christian Retoré, editor, *Logical Aspects of Computational Linguistics, Lecture Notes in Artificial Intelligence*, volume 1328, pages 329–345. https://doi.org/10.1007/BFb0052165

Miller, George A. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41. https://doi.org/10.1145/219717.219748

Ojeda, Almerindo. 1988. A linear precedence account of cross-serial dependencies. *Linguistics and Philosophy*, 11:457–492. https://doi.org/10.1007/BF00668683

Pagin, Peter and Dag Westerståhl. 2010. Pure quotation and general compositionality. *Linguistics and Philosophy*, 33:381–415. https://doi.org/10.1007/s10988-011-9083-8

Partee, Barbara. H., Alice ter Meulen, and Robert E. Wall. 1990. *Mathematical Methods in Linguistics*. Kluwer.

Pullum, Geoffrey K. and Gerald Gazdar. 1982. Natural languages and context-free languages. *Linguistics and Philosophy*, 4:471–504. https://doi.org/10.1007/BF00360802

Pullum, Geoffrey K. and Kyle Rawlins. 2007. Argument or no argument? *Linguistics and Philosophy*, 30:277–287. https://doi.org/10.1007/s10988-007-9013-y

Radzinski, Daniel. 1990. Unbounded syntactic copying in Mandarin Chinese. *Linguistics and Philosophy*, 13:113–127. https://doi.org/10.1007/BF00630518

Radzinski, Daniel. 1991. Chinese number-names, tree adjoining languages, and mild context-sensitivity. *Computational Linguistics*, 17:277–299.

Seki, Hiroyuki, Takashi Matsumura, Mamoru Fujii, and Tadao Kasami. 1991. On multiple context-free grammars. *Theoretical Computer Science*, 88:191–229. https://doi.org/10.1016/0304-3975(91)90374-B

Shieber, Stuart. 1985. Evidence against the context-freeness of natural language. *Linguistics and Philosophy*, 8:333–343. https://doi.org/10.1007/BF00630917

Stabler, Edward P. 2019. Three mathematical foundations for syntax. *Annual Review of Linguistics*, 5:243–260. https://doi.org/10.1146/annurev-linguistics-011415-040658

Wang, Shunjie and Shane Steinert-Threlkeld. 2023. Evaluating transformer's ability to learn mildly context-sensitive languages. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 271–283. https://doi.org/10.18653/v1/2023.blackboxnlp-1.21

Zhong, Zhi and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83.