Semilinearity in Old Georgian

Richard Luo

October 2, 2023

Background

After the context-freeness of natural language was argued against in Shieber (1985), there were different proposals for how to loosen the restrictions on CFGs so as to minimally categorize them.

Background

After the context-freeness of natural language was argued against in Shieber (1985), there were different proposals for how to loosen the restrictions on CFGs so as to minimally categorize them.

In response to this, Joshi (1985) put forward the idea of *mildly context-sensitive languages.*

Definitions

Definition

The class of **mildly context-sensitive languages** is characterized by the following properties:

- worst-case parsing complexity is polynomial, i.e. $O(n^k)$ for some $k \in \mathbb{N}$ where n is the sentence length/$\#$ of morphemes
- grammars can only capture a limited set of patterns of nested and crossed dependencies (e.g. Dutch coordination)
- context-free languages are a proper subset
- satisfy the *Constant Growth Property*

Definitions

Definition

A language L is of **constant growth** if there exists $c, c_0 > 0$ such that for any sentence $\alpha \in L$ with $|\alpha| \geq c_0$, there exists another sentence $\alpha' \in L$ satisfying $|\alpha| \leq |\alpha'| + c$.

Definitions

A closely related notion to constant growth is that of *semilinearity*:

Definitions

A closely related notion to constant growth is that of *semilinearity*:

Definition

Let $M \subseteq \mathbb{F}^n$ be a nontrivial subset of an n-dimensional vector space, where $n \in \mathbb{N}$. We say M is **linear** if for some $k \in \mathbb{N}$, there exist vectors $u^{(0)}, \ldots, u^{(k)} \in \mathbb{F}^n$ such that

$$M = \left\{ u^{(0)} + \sum_{i=1}^{k} n_i u^{(i)} \;\middle|\; n_i \in \mathbb{N} \text{ for } i = 1, \ldots, k \right\}.$$

Definitions

A closely related notion to constant growth is that of *semilinearity*:

Definition

Let $M \subseteq \mathbb{F}^n$ be a nontrivial subset of an n-dimensional vector space, where $n \in \mathbb{N}$. We say M is **linear** if for some $k \in \mathbb{N}$, there exist vectors $u^{(0)}, \ldots, u^{(k)} \in \mathbb{F}^n$ such that

$$M = \left\{ u^{(0)} + \sum_{i=1}^{k} n_i u^{(i)} \;\middle|\; n_i \in \mathbb{N} \text{ for } i = 1, \ldots, k \right\}.$$

Definition

We say M is **semilinear** if there are some linear $M_1, \ldots, M_k \subseteq \mathbb{F}^n$ such that $M = \bigcup_{i=1}^{k} M_i$.

More definitions

To see how semilinearity plays a role in studying formal languages, let us first introduce an important way to translate between them and vector spaces more generally:

Definition

Let Σ be an alphabet of size $n \in \mathbb{N}$, whose letters are enumerated $\{w_0, w_1, \ldots, w_{n-1}\}$. Also, let $e^{(0)}, \ldots, e^{(n-1)}$ denote the standard basis vectors of $\mathbb{R}^n$. The **Parikh mapping** $p_\Sigma : \Sigma^* \to \mathbb{N}^n$ is defined inductively as follows:

- $\varepsilon \mapsto \mathbf{0}$,
- $w_i \mapsto e^{(i)}$,
- for any two words $\alpha, \beta \in \Sigma^*$, we have $\alpha^\frown\beta \mapsto p_\Sigma(\alpha) + p_\Sigma(\beta)$.

More definitions

Definition

The **Parikh image** of the language $L \subseteq \Sigma^*$ is the set

$$p_\Sigma[L] := \{p_\Sigma(\alpha) \mid \alpha \in L\}.$$

More definitions

Definition

The **Parikh image** of the language $L \subseteq \Sigma^*$ is the set

$$p_\Sigma[L] := \{p_\Sigma(\alpha) \mid \alpha \in L\}.$$

Definition

If the Parikh image of L is semilinear, then L is said to be a **semilinear language.**

Some nice results

Theorem

A language $L \subseteq \Sigma^*$ is semilinear only if it is of constant growth.

Proof.

We will prove the case where L is linear. Let $|\Sigma| = n$, and suppose there are vectors $u^{(0)}, \dots, u^{(i-1)} \in \mathbb{N}^n$ such that

$$p_\Sigma[L] = \left\{ u^{(0)} + \sum_{i=1}^{k} n_i u^{(i)} \,\middle|\, n_i \in \mathbb{N} \text{ for } i = 1, \dots, k \right\}.$$

Let $c = \max_i \left\{ \sum_{j=1}^{n} u_j^{(i)} \right\}$. Then for any $\alpha \in p_\Sigma[L]$, it follows that there be some $\alpha' \in p_\Sigma[L]$ s.t. $\alpha \neq \alpha'$ and $|\alpha| \leq |\alpha'| + c$. $\square$

Some nice results

This next result is due to Parikh (1961):

Lemma

A language is semilinear if and only if it is letter-equivalent to a regular language.

Theorem

The Parikh image of a context-free language is semilinear. Equivalently, every context-free language has the same Parikh image as some regular language.

Some nice results

These two properties of semilinear sets will be central to our discussion of Old Georgian to follow:

Some nice results

These two properties of semilinear sets will be central to our discussion of Old Georgian to follow:

Theorem

For any $\alpha = (a_0, \ldots, a_m) \in \mathbb{R}^{m+1}$ *with* $a_m > 0$, *let* P_α *be the real polynomial of degree* m *corresponding to* α, $P_\alpha : x \mapsto \sum_{i=0}^{m} a_i x^i$ *for all* $x \in \mathbb{R}$. *Suppose* $M \subseteq \mathbb{N}^n$ *has the following properties:*

- *for any* $k \in \mathbb{N}^+$, *there exists* $\ell_2^{(k)}, \ldots, \ell_{n-1}^{(k)} \in \mathbb{N}$ *such that* $(k, P_\alpha(k), \ell_2^{(k)}, \ldots, \ell_{n-1}^{(k)}) \in M$,
- *the value* $P_\alpha(k)$ *provides an upper bound for the second component* ℓ_1 *of any tuple* $(k, \ell_1, \ldots, \ell_{n-1}) \in M$.

Then M *is not semilinear.*

Some nice results

These two properties of semilinear sets will be central to our discussion of Old Georgian to follow:

Theorem

For any $\alpha = (a_0, \ldots, a_m) \in \mathbb{R}^{m+1}$ *with* $a_m > 0$, *let* P_α *be the real polynomial of degree* m *corresponding to* α, $P_\alpha : x \mapsto \sum_{i=0}^{m} a_i x^i$ *for all* $x \in \mathbb{R}$. *Suppose* $M \subseteq \mathbb{N}^n$ *has the following properties:*

- *for any* $k \in \mathbb{N}^+$, *there exists* $\ell_2^{(k)}, \ldots, \ell_{n-1}^{(k)} \in \mathbb{N}$ *such that* $(k, P_\alpha(k), \ell_2^{(k)}, \ldots, \ell_{n-1}^{(k)}) \in M$,
- *the value* $P_\alpha(k)$ *provides an upper bound for the second component* ℓ_1 *of any tuple* $(k, \ell_1, \ldots, \ell_{n-1}) \in M$.

Then M *is not semilinear.*

Proposition

Let $M, N \in \mathbb{N}^n$ *be semilinear. Then* $M \cap N$ *is also semilinear.*

Stacking case suffixes in Old Georgian

Old Georgian is one of many languages that exhibits the phenomenon known as *Suffixaufnahme (lit. taking up of suffixes).* The OG grammar allows for multiple case(-number)-marking of nouns (Boeder 1995), each additional case marker being the result of some indirect case assignment.

Stacking case suffixes in Old Georgian

Basic form (prenominal):

[[[[Davit-is] galob-isa] muql-ta ama-t]
David-Gen singing-Gen verse-Pl(Gen) Art-Pl(Gen)
çartkuma-j]
recitation-Nom

'the recitation of the verses of the song of David'

Derived form (postnominal):

[saidumlo-j igi [sasupevel-isa m-is
mystery-Nom Art-Nom kingdom-Gen Art-Gen
[**mrt-isa-jsa-j**]]]
God-Gen-Gen-Nom

'the mystery of the kingdom of God'

Stacking case suffixes in Old Georgian

Multiple case stacking also appears to be a recursive operation:

[govel-i	igi	sisxl-i	[**saxl-isa-j**	m-is
all-Nom	Art-Nom	blood-Nom	**house-Gen-Nom**	Art-Gen

[**Saul-is-isa-j**]]]
Saul-Gen-Gen-Nom

'all the blood of the house of Saul'

Stacking case suffixes in Old Georgian

According to this observation, Michaelis and Kracht (1996) propose the following general form for complex nominative NPs, consisting of k stacked NPs where $k \in \mathbb{N}^+$:

$$\mathrm{N}_1 - \mathrm{Nom} \quad \mathrm{N}_2 - \mathrm{Gen} - \mathrm{Nom} \ldots \mathrm{N}_k - \mathrm{Gen}^{k-1} - \mathrm{Nom}$$

Stacking case suffixes in Old Georgian

According to this observation, Michaelis and Kracht (1996) propose the following general form for complex nominative NPs, consisting of k stacked NPs where $k \in \mathbb{N}^+$:

$$\text{N}_1 - \text{Nom} \quad \text{N}_2 - \text{Gen} - \text{Nom} \ldots \text{N}_k - \text{Gen}^{k-1} - \text{Nom}$$

From the above generalization, it follows that the number of genitive suffixes is bounded by the polynomial $\frac{k^2-k}{2}$.

Showing that Old Georgian is not semilinear

Here's the setup: take these lexical items from the OG alphabet Σ.

- w_0: some fixed noun (stem),
- w_1: genitive suffix,
- w_2: nominative suffix,
- w_3: genitive article,
- w_4: nominative article,
- w_5: some fixed intransitive verb

Showing that Old Georgian is not semilinear

Consider the linear (and hence semilinear) set

$$R = \left\{ e^{(4)} + e^{(5)} + \sum_{i=0}^{3} n_i e^{(i)} \mid n_i \in \mathbb{N} \text{ for } i = 0, 1, 2, 3 \right\}.$$

Showing that Old Georgian is not semilinear

Consider the linear (and hence semilinear) set

$$R = \left\{ e^{(4)} + e^{(5)} + \sum_{i=0}^{3} n_i e^{(i)} \mid n_i \in \mathbb{N} \text{ for } i = 0, 1, 2, 3 \right\}.$$

Its full pre-image under the Parikh mapping is the language

$$L_R := p_{\Sigma}^{-1}[R] = \{\alpha \in \Sigma^* \mid \text{ there is a } u \in R \text{ with } p_{\Sigma}(\alpha) = u\}.$$

Showing that Old Georgian is not semilinear

Consider the linear (and hence semilinear) set

$$R = \left\{ e^{(4)} + e^{(5)} + \sum_{i=0}^{3} n_i e^{(i)} \mid n_i \in \mathbb{N} \text{ for } i = 0, 1, 2, 3 \right\}.$$

Its full pre-image under the Parikh mapping is the language

$$L_R := p_\Sigma^{-1}[R] = \{\alpha \in \Sigma^* \mid \text{ there is a } u \in R \text{ with } p_\Sigma(\alpha) = u\}.$$

This language consists of all strings of words (in no particular order) with only one appearance of $w^{(4)}$ and $w^{(5)}$, and arbitrarily many appearances of $w^{(0)}, w^{(1)}, w^{(2)}, w^{(3)}$.

Showing that Old Georgian is not semilinear

To restrict only to sentences that are grammatical in Old Georgian, we take its intersection with the OG language $L_G \subseteq \Sigma^*$:

$$L_M := L_G \cap L_R.$$

Showing that Old Georgian is not semilinear

To restrict only to sentences that are grammatical in Old Georgian, we take its intersection with the OG language $L_G \subseteq \Sigma^*$:

$$L_M := L_G \cap L_R.$$

The Parikh mapping respects set intersection:

$$\begin{aligned} M &:= p_\Sigma(L_M) \\ &= p_\Sigma(L_G) \cap p_\Sigma(L_R) \\ &= p_\Sigma(L_G) \cap R. \end{aligned}$$

Showing that Old Georgian is not semilinear

What do we know about M? As we saw earlier, if $k \in \mathbb{N}^+$ is number of stacked nouns within the complex NP, the number of genitive suffixes that appear cannot exceed $\frac{k^2-k}{2}$. Thus, given $m_0 = k$, we obtain an upper bound on m_1 for any vector $(m_0, m_1, \ldots, m_5) \in M$, which counts appearances of $w^{(1)}$.

Showing that Old Georgian is not semilinear

What do we know about M? As we saw earlier, if $k \in \mathbb{N}^+$ is number of stacked nouns within the complex NP, the number of genitive suffixes that appear cannot exceed $\frac{k^2-k}{2}$. Thus, given $m_0 = k$, we obtain an upper bound on m_1 for any vector $(m_0, m_1, \ldots, m_5) \in M$, which counts appearances of $w^{(1)}$.

Furthermore, we assume there exist $m_2^{(k)}, m_3^{(k)} \in \mathbb{N}$ such that

$$(k, (k^2 - k)/2, m_2^{(k)}, m_3^{(k)}, 1, 1) \in M.$$

Showing that Old Georgian is not semilinear

What do we know about M? As we saw earlier, if $k \in \mathbb{N}^+$ is number of stacked nouns within the complex NP, the number of genitive suffixes that appear cannot exceed $\frac{k^2-k}{2}$. Thus, given $m_0 = k$, we obtain an upper bound on m_1 for any vector $(m_0, m_1, \ldots, m_5) \in M$, which counts appearances of $w^{(1)}$.

Furthermore, we assume there exist $m_2^{(k)}, m_3^{(k)} \in \mathbb{N}$ such that

$$(k, (k^2 - k)/2, m_2^{(k)}, m_3^{(k)}, 1, 1) \in M.$$

Thus, by the theorem mentioned before, we conclude that M is not semilinear.

Showing that Old Georgian is not semilinear

But recall that $M = p_\Sigma(L_G) \cap R$, and R is linear!

Showing that Old Georgian is not semilinear

But recall that $M = p_\Sigma(L_G) \cap R$, and R is linear!

Since semilinearity is closed under intersection, it follows that $p_\Sigma(L_G)$ is not semilinear, hence Old Georgian L_G is not a semilinear language.

Showing that Old Georgian is not semilinear

But recall that $M = p_{\Sigma}(L_G) \cap R$, and R is linear!

Since semilinearity is closed under intersection, it follows that $p_{\Sigma}(L_G)$ is not semilinear, hence Old Georgian L_G is not a semilinear language.

Do you notice any potential problems with this argument?

Refuting the non-semilinear claim

Bhatt and Joshi (2003) provide two reasons to argue against Michaelis and Kracht (1996) that Old Georgian is not semilinear, consistent with Boeder (1995)'s analysis of *Suffixaufnahme*:

- two types of *Suffixaufnahme*
- haplology induces morphological restrictions

Refuting the non-semilinear claim

According to Border (1995), there are two types of *Suffixaufnahme*, interactions between the two of which are not fully attested:

Refuting the non-semilinear claim

According to Border (1995), there are two types of *Suffixaufnahme*, interactions between the two of which are not fully attested:

Multiple Suffixaufnahme

$$\mathsf{N}_1 - \mathsf{Nom} \quad \mathsf{N}_2 - \mathsf{Nom} \ldots \mathsf{N}_k - \mathsf{Gen}^{k-1} - \mathsf{Nom}$$

Refuting the non-semilinear claim

According to Border (1995), there are two types of *Suffixaufnahme*, interactions between the two of which are not fully attested:

Multiple Suffixaufnahme

$$\mathsf{N}_1 - \mathsf{Nom} \quad \mathsf{N}_2 - \mathsf{Nom} \ldots \mathsf{N}_k - \mathsf{Gen}^{k-1} - \mathsf{Nom}$$

Simple Suffixaufnahme

$$\mathsf{N}_1 - \mathsf{Nom} \quad \mathsf{N}_2 - \mathsf{Gen} - \mathsf{Nom} \ldots \mathsf{N}_k - \mathsf{Gen} - \mathsf{Nom}$$

Refuting the non-semilinear claim

When these both take place, we could in principle get Michaelis and Kracht (1996)'s recursive formulation of case-marking in complex NPs:

$$\mathsf{N}_1 - \mathsf{Nom} \quad \mathsf{N}_2 - \mathsf{Gen} - \mathsf{Nom} \ldots \mathsf{N}_k - \mathsf{Gen}^{k-1} - \mathsf{Nom}$$

However, this requires multiple steps of simple *Suffixaufnahme*, which might be impossible since it is expected to only apply at the top layer (in particular to assign nominative case).

Refuting the non-semilinear claim

The general pattern would, in fact, look something more like:

$\mathsf{N}_1\text{—Nom} \quad \mathsf{N}_2\text{—Gen—Nom} \quad \mathsf{N}_3\text{—Gen—Nom} \ldots \mathsf{N}_k\text{—Gen}^{k-1}\text{—Nom}$

If so, the number of case suffixes still obeys constant growth!

Refuting the non-semilinear claim

There's also the issue of *haplology*, the process by which a whole syllable is deleted before or after a phonetically similar or identical syllable. Bhatt and Joshi (2003) observe that haplology is obligatory for plural genitive markers, but optional for singular:

Refuting the non-semilinear claim

There's also the issue of *haplology*, the process by which a whole syllable is deleted before or after a phonetically similar or identical syllable. Bhatt and Joshi (2003) observe that haplology is obligatory for plural genitive markers, but optional for singular:

z-isa	kac-isa-jsa
son-Gen	man-Gen-Gen

'the son of man'

z-isa	**kac-isa-∅**
son-Gen	**man-Gen**

'the son of man'

Refuting the non-semilinear claim

Obligatory deletion of repeated plural genitive marker:

*kar-ta	kalak-ta-ta
door-Pl(Obl)	city-Pl(Gen)-Pl(Gen)

Intended: 'the gates of the cities'

kar-ta	**kalak-ta-∅**
door-Pl(Obl)	**city-Pl(Gen)**

'the gates of the cities'

Refuting the non-semilinear claim

In light of this pattern, there have also been no instances of three stacked genitive suffixes in Old Georgian according to data from Boeder (1995). This can be reasonably explained as being a morphological constraint due to haplology.

Refuting the non-semilinear claim

In light of this pattern, there have also been no instances of three stacked genitive suffixes in Old Georgian according to data from Boeder (1995). This can be reasonably explained as being a morphological constraint due to haplology.

Therefore, even if the non-constant growth pattern by Michaelis and Kracht (1996) were permitted, it would be reduced to at most three suffixes per subsequent stacked noun, which is of CG:

$$\mathsf{N}_1 - \mathsf{Nom} \quad \mathsf{N}_2 - \mathsf{Gen} - \mathsf{Nom} \quad \mathsf{N}_3 - \mathsf{Gen} - \mathsf{Gen} - \mathsf{Nom} \ldots$$

$$\mathsf{N}_k - \mathsf{Gen} - \mathsf{Gen} - \mathsf{Nom}$$

Thanks for listening!