

# Mathematical Methods in Linguistics

## Prerequisites

- sets (notation)
- strings (basic notation)

Formal definition and proof of the normal form theorem

The previous two sections introduced negative  $n$ -gram grammars at great length and showed a basic normal form theorem: for every grammar with  $n$ -grams of mixed length, there is an equivalent grammar where all  $n$ -grams have the same length. The presentation was deliberately informal to focus on intuitions rather than mathematical rigor. This unit is very different. It gives the definitions in a mathematical format, rigorously states the normal form theorem, and states the proof of the theorem in a more standard mathematical style.

I admit that this might be a lot to take in for the newbie, but it is important for you to learn how to read mathematical notation. It really makes things a lot easier in the long run. Once you feel more comfortable with mathematical notation, I suggest that you come back to this unit and contrast it to the two preceding ones. Which one gives you more information in a short amount of time?

If you're suffering an acute case of symbol shock, don't worry. We will continue at a leisurely pace, with optional formal sections sprinkled in to give a succinct summary of the more informal sections.

## 1 Formal definition of negative grammars

An **alphabet** is a finite set of symbols.

---

**Definition 1.** Let  $\Sigma$  be some alphabet, and  $\Sigma_E$  its extension with a edge marker symbols  $\bowtie, \bowtie \notin \Sigma$ . An  $n$ -gram over  $\Sigma_E$  is an element of  $\Sigma_E^n$  ( $n \geq 1$ ). A **negative  $n$ -gram grammar**  $G$  over alphabet  $\Sigma$  is a finite set of  $n$ -grams over  $\Sigma_E$ . A string  $s$  over  $\Sigma$  is well-formed with respect to  $G$  iff there are no  $u, v$  over  $\Sigma_E$  and no  $g \in G$  such that  $\bowtie^{n-1} \cdot s \cdot \bowtie^{n-1} = u \cdot g \cdot v$ . The **language of**  $G$ , denoted  $L(G)$ , contains all strings that are well-formed with respect to  $G$ , and only those.

---

**Example 1** Suppose  $\Sigma := \{C, V\}$ , where  $C$  represents consonants and  $V$  vowels. One string over  $\Sigma$  is  $CVCVCV$ , an instance of a very simple CV-syllable template. Assume  $G$  contains  $CC$  and  $VC$  and let's see if the string  $CVCVCV$  is well-formed with respect to  $G$ . The bigram  $CC$  is not a problem since there are no strings  $u$  and  $v$  such that  $\bowtie CC \bowtie = u \cdot CC \cdot v$ , which means that  $CVCVCV$  does not contain the forbidden

bigram  $CC$ . But clearly  $\bowtie CVCV\bowtie = \bowtie C \cdot VC \cdot V\bowtie$ . So  $VC$  is a component of  $CVCV$ , and as a result the string is ruled out by  $G$ .

---

**Definition 2.** A **mixed negative  $n$ -gram grammar**  $G$  is a finite set of strings over  $\Sigma$  such that  $n$  is the length of the longest string in  $G$ . A negative  $n$ -gram grammar that is not mixed is called **strict**.

---

## 2 Normal form theorem

**Theorem 3.** For every mixed negative  $n$ -gram grammar  $G$ , there is a strict negative  $n$ -gram grammar  $G'$  such that  $L(G) = L(G')$ .

*Proof.* Let  $G' := \{u \cdot g \cdot v \mid g \in G, u, v \in \Sigma^*, \text{ and the length of } u \cdot g \cdot v \text{ is } n\}$ . Suppose  $s \notin L(G)$ . Then there must be some  $g \in G$  and  $u = u_1 \cdot u_2$  and  $v = v_1 \cdot v_2$  over  $\Sigma$  such that  $\bowtie^{n-1} \cdot s \cdot \bowtie^{n-1} = u \cdot g \cdot v$ . But then  $\bowtie^{n-1} \cdot s \cdot \bowtie^{n-1} = u_1 \cdot u_2 \cdot g \cdot v_1 \cdot v_2$ . As the length of  $\bowtie^{n-1} \cdot s \cdot \bowtie^{n-1}$  exceeds  $n$ , it holds that  $u_2 \cdot g \cdot v_1 \in G'$  for some choice of  $u_2$  and  $v_1$ . But then  $s \notin L(G')$ .

In the other direction, suppose  $s \notin L(G')$ . Then there is some  $g \in G'$  such that  $\bowtie^{n-1} \cdot s \cdot \bowtie^{n-1} = u \cdot g \cdot v$ . But then there must  $u', g'$  and  $v'$  over  $\Sigma$  such that  $g = u' \cdot g' \cdot v'$  and  $g' \in G$ . It follows that  $s \notin L(G)$ .

And there you have it. All the ground we've covered in dozens of pages so far, condensed into less than one page. That's the power of math.