Positive *n*-gram grammars (Solutions)

Exercise 1.
For each one of the following phenomena, write a negative *n*-gram grammar that handles it correctly. For some of them, you have to rephrase the phenomenon as a phonotactic constraint first.

1. **intervocalic voicing**: voiceless fricatives (assume *s* and *f*) may not occur between vowels (assume *a*, *i*, *u*)
2. **local assimilation**: *n* must be *m* before *b* or *p*
3. **local dissimilation**: *rVr* becomes *lVr*, where *V* is *a*, *i*, or *u*
4. **penultimate stress**: in words with at least two syllables, stress falls on the last but one syllable (assume that words are strings of stress syllables ($\acute{\sigma}$) and unstressed syllables (σ))

Solution

1. **intervocalic voicing**: asa, asi, asu, afa, afi, afu, isa, isi, isu, ifi, ifa, ifu, usa, usi, usu, ufa, ufu, ufi
2. **local assimilation**: nb, np
3. **local dissimilation**: rar, rir, rur
4. **penultimate stress** $\rtimes \sigma \ltimes, \sigma \acute{\sigma} \ltimes, \sigma \sigma \ltimes, \acute{\sigma} \sigma \sigma$

Explanation

1. We have to forbid trigrams of the form *xyz* where *x* and *z* are vowels and *y* is a voiceless fricative. The exercise tells us that *x* and *z* can be *a*, *i*, or *u*, and *y* can be *s* or *f*. After carrying out all possible substitutions of *x*, *y*, and *z*, we get the list of forbidden trigrams above.
2. If *n* must be *m* before *b* or *p*, this means that we can never have an *n* followed by *b* or *p* (because then we would have a case where *n* failed to turn into *m*). Hence we forbid the bigrams *nb* and *np*.
3. The exercise describes local dissimilation as a process where *r* changes to *l* if it is followed by *Vr*. In terms of phonotactics, this means that we cannot have *r* followed by *Vr*. Replacing *V* with all possible choices for a vowel, we get the forbidden trigrams *rar*, *rir*, and *rur*.
4. This one is tricky because we have to consider multiple cases. If a word has just one syllable, then it is not subject to the penultimate stress rule (which only applies to words with at least two syllables), and hence stress falls on the last syllable in this case. In other words, monosyllabic words must be of the form $\acute{\sigma}$, whereas σ is not allowed. If a word has exactly two syllables, then it must be of the form $\acute{\sigma}\sigma$ because the other option $\sigma\acute{\sigma}$ would violate the penultimate stress rule. And if a word has three or more syllables, then

it must be of the form $\sigma^+\acute{\sigma}\sigma$. Based on this, we can deduce what must be forbidden: First, we do not want to allow words like σ, so our grammar must contain the forbidden trigram $\rtimes\sigma\ltimes$. We also do not want to allow any word with at least two syllables where stress falls on the last syllable, and thus we add the forbidden trigram $\sigma\acute{\sigma}\ltimes$. We do not want to allow any word that ends with two unstressed syllables, which we capture with the forbidden trigram $\sigma\sigma\ltimes$. The combination of $\sigma\acute{\sigma}\ltimes$ and $\sigma\sigma\ltimes$ guarantees that if a word has at least two syllables, the last but one is stressed. But that's not quite enough, we also have to ensure that this is the only syllable that is stressed. That is the same as saying that we do not to allow any word where the stressed syllable is followed by at least two unstressed syllables, and thus we also forbid the trigram $\acute{\sigma}\sigma\sigma$. And that's it, nothing else is required.

Exercise 2.

For each one of the following *n*-grams, say how large it is depending on what one chooses as the basic symbols that *n*-grams are built from. Possible choices for building blocks are typed characters, morphemes, or words. Not all choice may be appropriate in each case.

1. *de-*
2. *mpi*
3. *John likes Mary*

Solution

1. *de-*: It could be a unigram consisting just of the morpheme *de-*, or a trigram that consists of the characters *d*, *e*, and a hyphen.
2. *mpi*: The only reasonable treatment here is as a trigram consisting of the characters *m*, *p*, and *i*.
3. *John likes Mary*: This could be a trigram that consists of the words *John*, *likes*, and *Mary*, or a 4-gram that consists of the morphemes *John*, *like*, *s*, and *Mary*, or a 15-gram that consists of a long sequence of letters and spaces.

Exercise 3.

Write both a positive and a negative grammar that each allow only strings of the form *ab*, *abab*, *ababab*, *abababab*, and so on (assume that all symbols are either *a* or *b*). Is one of the two grammars more succinct or general than the other? What if the set of symbols is larger, e.g. *a*, *b*, *c*, and *d*?

Solution

- **positive grammar**: $\rtimes a, b\ltimes, ab, ba$
- **negative grammar**: $\rtimes b, a\ltimes, aa, bb, \rtimes\ltimes$

The two grammars are very similar in size, the only difference is that the negative grammar also has to block the empty string. However, when the set of symbols gets larger, the positive grammar stays the same but the negative grammar will need more bigrams to rule out the illicit combinations.

Explanation

The general pattern is that strings must start with a, must end with b, and that a and b must alternate. The positive grammar states this very directly with its bigrams $\rtimes a$, ab, ba, and $b\ltimes$. The negative grammar, on the other hand, captures this more indirectly. For example: instead of “start with a”, the negative grammars says “start with a symbol” ($\rtimes\ltimes$) and “do not start with b” ($\rtimes b$). But if there’s more symbols than just a and b, the negative grammar also has to say “do not start with c” ($\rtimes c$), “do not start with d” ($\rtimes d$), and so on. As the number of symbols increases, more and more options need to be blocked.

Quite generally, a negative grammar will be very large when the majority of logically possible combinations is impossible. In this case, a positive grammar is smaller. But on the flip side, if most combinations are allowed and only a small number need to be ruled out, the negative grammar will be smaller.

Exercise 4.

For each one of the following phenomena, write a positive n-gram grammar that handles it correctly. For some of them, you have to rephrase the phenomenon as a phonotactic constraint first.

- **intervocalic voicing**: voiceless fricatives (assume s and f) may not occur between vowels (assume a, i, u)
- **local assimilation**: n must be m before b or p
- **local disimilation**: rVr becomes lVr, where V is a, i, or u
- **penultimate stress**: in words with at least two syllables, stress falls on the last but one syllable (assume that words are strings of stress syllables ($\acute{\sigma}$) and unstressed syllables (σ))

Once you’re done, contrast the positive grammars against the negative ones from an earlier exercise. Can you identify some general guidelines for when a positive grammar is preferable to a negative one?

Solution

This exercise requires a lot more assumptions than the one for negative grammars. The problem is that it is not enough to know how intervocalic voicing works, we also have to know what the rest of the language looks like. Consider the case of intervocalic voicing. With the negative grammar, it was enough to say that s and f may not occur between the vowels a, i, and u. With the positive grammar, we instead have to allow for every possible trigram except the ones where s or f occurs between a, i and u. But we do not actually know what the set of all possible trigrams is because the exercise does not specify the alphabet. If the alphabet contains only *a*, *i*, *u*, *s*, *f*, and *k*, then the set of possible trigrams is much smaller compared to an alphabet that also contains all the other consonants of English.

For this reason, each answer must always specify the assumed alphabet. And ideally, this alphabet will be small to reduce the number of *n*-grams that need to be written down. Even then, though, these grammars will be very large. They contain all possible *n*-grams except the ones that were listed in the negative grammar.

Exercise 5.

English allows for *nature*, *natural*, *naturalize*, *denaturalize*, *naturalization*, and *denaturalization*, but not *denature* or any other misordered forms like *naturizalation* (actually there is a technical term *denature*, but it is not very common and we'll exclude it here to keep the exercise simple). Write a grammar that generates all the well-formed forms but none of the ill-formed ones. It is up to you whether you want to use a positive or a negative grammar. If you use a negative grammar, it can be in the mixed format, with *n*-grams of varying lengths.

Solution

A negative mixed grammar is the easiest option here. As our alphabet, we assume the morphemes *de-*, *nature*, *-al*, *-ize*, and *-ation*. The negative grammar then contains the following *n*-grams:

1. You must start with *de-* or *nature*: ⋊ -al, ⋊ -ize, ⋊ -ation
2. *de-* can only be followed by *nature*: de- de-, de- -al, de -ize, de -ation, de ⋉
3. *nature* can only be followed by *-al* or the end of the word: nature de-, nature nature, nature -ize, nature -ation
4. *-al* can only be followed by *-ize* or the end of the word: -al de-, -al nature, -al -al, -al -ation
5. *-ize* can only be followed by *-ation* or the end of the word: -ize -de, -ize nature, -ize -al, -ize -ize
6. *-ation* can only be followed by the end of the word: -ation -de, -ation nature, -ation -al, -ation -ize, -ation -ation
7. Do not end in *denature* or *denatural*: de- nature ⋉, de- nature -al ⋉