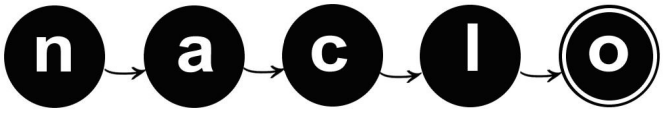
(H) Sequitur (1/2) [5 Points]

Space is valuable, so it's often desirable to compress data — that is, to use less space to convey the same information. One common data compression strategy is to identify repeated patterns within the data and somehow consolidate these repetitions.[1] First, let's look at Sequitur, a fast compression algorithm that uses the repeated pattern strategy. The table below shows Sequitur running on the input "abcdbcabcd", with its output at the very bottom:

Step Number	**Column 1**	**Column 2**
1)	a	S = a
2)	ab	S = ab
3)	abc	S = abc
4)	**(a)**	S = abcd
5)	abcdb	**(b)**
6)	abcdbc	S = abcdbc
	aXdX	S = aXdX X = bc
7)	**(c)**	**(d)** **(e)**
8)	aXdXab	**(f)** X = bc
9)	aXdXabc	S = aXdXabc X = bc
	aXdXaX	S = aXdXaX **(g)**
	(h)	**(i)** X = bc Y = aX
10)	**(j)**	**(k)** X = bc Y = aX
	ZXZ	S = ZXZ X = bc Y = aX Z = Yd
S = ZXZ; X = bc; Y = aX; Z = Yd		

[1] It just so happens that identifying repeated patterns in a human language can tell us a lot about how that language works. If you want to know more about how linguists (and especially computational linguists) use compression in language technologies, try problem (I), Non Sequitur, also from NACLO 2021, Round 1. Be aware, however, that solving problem (I) will not give you any advantage in solving this problem.



(H) Sequitur (2/2)

H1. Fill in the blanks (marked with bolded letters) in the table on the previous page.

Next, take a look at Byte Pair Encoding (BPE), a widely used compression algorithm that also uses the repeated pattern strategy. Like in the previous example, BPE is running with an input of "abcdbcabcd", and its output is shown at the bottom of the table:

Step Number	**Column 1**	**Column 2**
1)	abcdbcabcd	S = abcdbcabcd
2)	**(a)**	**(b)** X = bc
3)	YdXYd	S = YdXYd X = bc **(c)**
4)	ZXZ	S = ZXZ X = bc Y = aX Z = Yd
S = ZXZ; X = bc; Y = aX; Z = Yd		

H2. Fill in the blanks (marked with bolded letters) in the table above.

As you can see, for the input "abcdbcabcd", Sequitur and BPE produce the same output! But this isn't always the case.

H3. For each of the following inputs, say whether Sequitur and BPE give the same or different outputs. Circle the correct response:

a) abcabdbcbc *Same* *Different*

b) abbcaddca *Same* *Different*

c) bacbcbabacbcba *Same* *Different*

d) ccdbccdbccacc *Same* *Different*

e) ccdbccdbccaccacca *Same* *Different*