CS 172 Handout

In class we discussed the language model and specifically the unigram and bigram models. Assume you have parsed a document collection and extracted the following probabilities for bi-grams.

Eat on	.16	Eat Thai	.03
Eat some	.06	Eat breakfast	.03
Eat lunch	.06	Eat in	.02
Eat dinner	.05	Eat Chinese	.02
Eat at	.04	Eat Mexican	.02
Eat a	.04	Eat tomorrow	.01
Eat Indian	.04	Eat dessert	.007
Eat today	.03	Eat British	.001
<start> I	.25	Want some	.04
<start> I'd	.06	Want Thai	.01
<start> Tell	.04	To eat	.26
<start> I'm	.02	To have	.14
I want	.32	To spend	.09
I would	.29	To be	.02
I don't	.08	British food	.60
I have	.04	British restaurant	.15
Want to	.65	British cuisine	.01
Want a	.05	British lunch	.01

Use the table above to extract the probabilities of the bi-grams:

P(I want to eat British food) = P(I|<start>) * P(want|I) * P(to|want) * P(eat|to)
* P(British|eat) * P(food|British)
=

What about the following example? How could we handle un-seen bi-grams?

P(I want to eat British but not Mexican) = ??