メモ：文字列の確率と EOS

文字列の終わりを表す特殊文字 EOS を考えることは, 文字列がどんな風に終わりやすいかという言語の性質を表せるだけでなく, 実は, 確率モデルとしても不可欠な要素です. 図 2.9 に示したように, EOS を考えると, “”(空文字列), “a”, “aa”, “ab”, … といったすべての文字列は先頭から順に文字を選択し, EOS が出たときそこで止まった結果として表すことができます. あらゆる文字列はこうした選択の結果に対応していますから, それらの確率の総和は必ず 1 になります. いっぽう, EOS がないと “aaaaa…” といった, いくらでも長い文字列にも一定の確率を割り当てることができ, それらが無限にありますから, 確率の総和は容易に 1 を超えてしまいます. これは, 文字列の確率モデルとしては不適切です.

なお, 再帰的な選択によって文字列を生成できることから, 次の文字の選択肢を可視化して, その幅を確率に比例させれば, $[0,1)$ の範囲の実数を次々と指示することで, 効率的に文字を入力することができます [*20]. 図 2.10 に示した MacKay らによる Dasher [13] [*21] は, このことを利用した入力システムで, もし病気や障害で指が動かない場合でも, まぶたの上下や, 呼吸による腹の上下といった 1 次元のわずかな信号さえあれば, 言葉を発することができます.

これらは文字列に限らず, 3 章の単語列の場合でもすべて同様に成り立ちます.

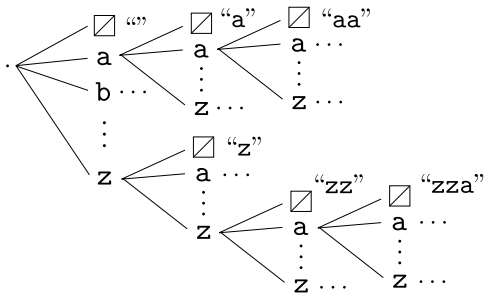

図 2.9: 文字列全体の空間を表す樹形図. ☒で EOS を表しています.

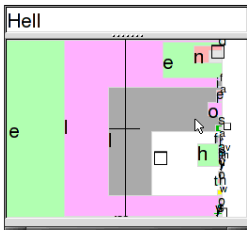

図 2.10: 入力システム Dasher で “Hello” を入力している様子. 1 次元の選択を次々に行うだけで, 文字列を入力することができます.

*20　これは情報理論では, **算術符号**とよばれる符号化を行っていることに相当しています.