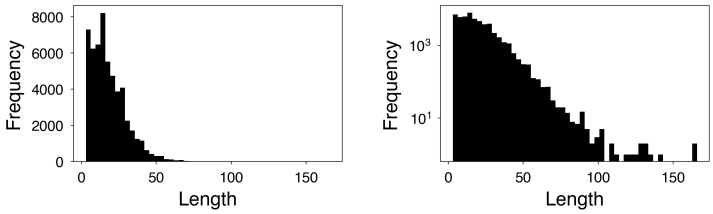
類似度	文	類似度	文
1.0000	だが、新しい後期高齢者医療制度では、介護..	1.0000	再生可能なファイル形式は、映像が MPEG..
0.8929	健康保険や介護保険、厚生年金、雇用保険、..	0.9054	無線 LAN セキュリティは 64/128bit の W..
0.8705	国は患者が混合診療を受けた場合、「一体化..	0.8790	その他の機能は地上デジタル/BS デジタル/..
0.8597	労働保険は、法人個人を問わず労働者を 1 人..	0.8780	ネットワーク機能は 10/100/1000BASE-T..
0.8313	また、短期入所や通所を受け入れる福祉施設..	0.8731	録音形式はリニア PCM で 16bit/44.1kHz..
0.8212	「住宅ローン控除」は、国内で一定の居住用..	0.8708	基本仕様は MP3/WMA/AAC 再生。
0.8109	全体で５％アップと同水準だが、保険制度の..	0.8627	その他の機能は、IEEE802.11b/g/n 対応..
0.8107	この免除の手続きをするだけで、保険料を払..	0.8588	入出力端子には HDMI/コンポジットビデオ..
0.8098	「年金制度は世代間扶養の仕組みである」→..	0.8486	同サービスは、i モード/EZweb/Yahoo!ケ..
0.8085	また、介護保険は対象外となっています。	0.8485	CG-BARPROG-X コレガは、WAN/LAN..
0.8013	語学学校は特定商取引法の指定業務で、受講..	0.8417	対応 OS は、WindowsXP(SP2/SP3)/Vi..
0.7946	機構や文部科学省によると、新制度は、悪質..	0.8383	その他の機能は、10BASE-T/100BASE-T..
0.7937	連合はほかに「中低所得者層の所得税減税」..	0.8377	「morawin[モーラウィン]forS!ミュージック..
0.7777	農水省案では、減反に加わる農家には生産量..	0.8368	Blu-ray ディスク作成においては、1080i/7..
0.7740	厚生労働省は２６日、サービス事業者に支払..	0.8339	対応ゲスト OS は OracleEnterpriseLinux..

図 4.5: Leipzig コーパスの日本語ニュースの文について, uSIF による文ベクトルを用いて計算した類似文. 数字はコサイン類似度を表します. 自分自身との類似度は 1 ですが, それ以外も意味的に類似した文が, 簡単な計算で検索できることがわかります. 単語ベクトルは, `text8` で計算した 100 次元のものを用いました.

ノート：文の長さの統計モデル

文の長さは, どのように分布しているのでしょうか. 図 4.6 は, 単語で数えた Brown コーパスの文の長さの頻度分布と, その縦軸を対数で表したものです. (a) では見えませんが, (b) を見ると, 非常に長い文があるためにこの分布は裾が長く, 対数スケールでは直線的に落ちる分布となっていることがわかります [*14]. 文の長さの確率分布は, 深層学習を使う場合でも無限の繰り返しが発生しうる文の生成を制御するために必要ですし, 子供の発達段階でみられる文の複雑さを定量化する[127]ことなどにも有効です.



(a) 横軸が文長, 縦軸が頻度を表します. (b) 縦軸を対数軸で表した場合.

図 4.6: Brown コーパスでの文の長さ (単語数) の分布.

*14 これは, 複数の従属節が接続詞などで繋がって長い文となることがあるためです[128].

文の長さの確率分布は, 1887 年の *Science* 誌[129]でこれが著者によって異なり, 著者の識別に有効なことが発見されて以来, ガンマ分布や負の二項分布[130], 対数正規分布[131]などでモデル化されてきました. 特に, 文の長さ n は対数をとると正規分布 (式(4.33)) に従うように一見みえますが, 正確にはそうではなく, Sichel は 1974 年に, 次式で表される Sichel 分布でよく当てはめられることを示しました[132].

$$p(n|\theta,\alpha,\gamma) = \frac{\sqrt{1-\theta}^{\,\gamma}}{K_\gamma(\alpha\sqrt{1-\theta})}\frac{(\alpha\theta/2)^n}{n!}K_{n+\gamma}(\alpha) \tag{4.17}$$

ここで $K_\gamma(x)$ は次数 γ の第二種変形ベッセル関数 [*15] で, $0<\theta<1$, $\alpha>0$, $\gamma\in\mathbb{R}$ は分布のパラメータです. Sichel 分布は, ポアソン分布

$$\mathrm{Po}(n|\lambda) = \frac{\lambda^n e^{-\lambda}}{n!} \tag{4.18}$$

の期待値 λ が逆ガンマ分布に従うとした場合の複合分布で, 特別な場合としてポアソン分布や負の二項分布, Yule 分布などを含んでいます. 特に, $\gamma=-1/2$ の場合をポアソン逆ガンマ分布といい, 文長のモデル化に用いられました. 図 4.7 に, この分布を文字数で測った京大コーパスの文長の分布に当てはめた例を示しました [*16]. 裾の部分も含め, 非常によい近似となっていることがわかります. 最近では, 文長を 1 次元のランダムウォークが原点に戻るまでの再帰時間とみたモデルが, Sichel 分布より χ^2 検定でさらによく当てはまることが示されています[133].

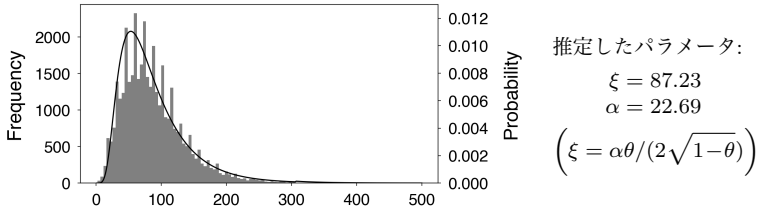


図 4.7: 京大コーパスの文長の分布とポアソン逆ガンマ分布による当てはめ. 分布の凹凸は, 全角文字の長さを 2 文字として数えているためです.

∗15　Python では, `scipy.special.kv` で計算することができます.