

Definition

Sei $a \in \Sigma^r$ ein Text (mit $r \geq 2$) und $\kappa_1(a), \dots, \kappa_{r-1}(a)$ die Folge der Autokoinzidenzindizes. Dann heißt der Mittelwert

$$\varphi(a) := [\kappa_1(a) + \dots + \kappa_{r-1}(a)] / [r - 1]$$

der (globale) **Koinzidenzindex** von a .

Dadurch ist also eine Abbildung

$$\varphi : \Sigma^{(\geq 2)} \rightarrow \mathbb{Q}$$

definiert.

Andere Beschreibung

Wie oft findet man, wenn man zwei beliebige Stellen des Textes a herauspicks, ein »Zwillingspaar«, also zweimal den gleichen Buchstaben $s \in \Sigma$? (= eine »Koinzidenz«)

Sei $h_s = \#\{j \mid a_j = s\}$ die Häufigkeit von s in a . Dann ist die Antwort:

$$h_s \cdot (h_s - 1) / 2 \text{ -mal.}$$

Insgesamt ist die Anzahl der Zwillingspaare also

$$\sum_{s \in \Sigma} \frac{h_s(h_s - 1)}{2} = \frac{1}{2} \cdot \sum_{s \in \Sigma} h_s^2 - \frac{1}{2} \cdot \sum_{s \in \Sigma} h_s = \frac{1}{2} \cdot \sum_{s \in \Sigma} h_s^2 - \frac{r}{2}.$$

Man kann diese Koinzidenzen (= Zwillingspaare) auch anders zählen:

Sei dazu z_q die Anzahl der bereits gefundenen Koinzidenzen im Abstand q für $q = 1, \dots, r-1$, initialisiert mit $z_q := 0$.

Für $i = 0, \dots, r-2$

für $j = i+1, \dots, r-1$

falls $a_i = a_j$

inkrementiere z_{j-i}

inkrementiere z_{r+i-j}

[Zählschleife über den Text a]

[Zählschleife über den Resttext]

[Koinzidenz gefunden]

[im Abstand $j-i$]

[und im Abstand $r+i-j$]

Genauso wurde übrigens in dem [Perl-Programm](#) gezählt.

Am Ende der Schleife enthalten z_1, \dots, z_{r-1} Werte, für die gilt:

Hilfssatz. (i) $z_1 + \dots + z_{r-1} = \sum_{s \in \Sigma} h_s \cdot (h_s - 1)$.

(ii) $\kappa_q(a) = z_q / r$ für $q = 1, \dots, r-1$.

Beweis. (i) Alle Koinzidenzen sind doppelt gezählt.

(ii) $\kappa_q(a) = (1/r) \cdot \#\{j \mid a_{j+q} = a_j\}$ nach Definition (Indizes mod r). ♦

Satz (Kappa-Phi-Theorem) *Der Koinzidenzindex eines Textes $a \in \Sigma^*$ (der Länge $r \geq 2$) ist der Anteil der Koinzidenzen unter allen Buchstabenpaaren von a .*

Beweis.

$$\begin{aligned} \varphi(a) &= \frac{\kappa_1(a) + \dots + \kappa_{r-1}(a)}{r-1} = \frac{z_1 + \dots + z_{r-1}}{r \cdot (r-1)} \\ &= \frac{\sum_{s \in \Sigma} h_s \cdot (h_s - 1)}{r \cdot (r-1)} = \frac{\sum_{s \in \Sigma} \frac{h_s(h_s - 1)}{2}}{\frac{r \cdot (r-1)}{2}}; \end{aligned}$$

im Zähler steht jetzt die Gesamtzahl der Koinzidenzen, im Nenner die Gesamtzahl aller Buchstabenpaare. ♦

Korollar 1.

$$\varphi(a) = \frac{r}{r-1} \cdot \sum_{s \in \Sigma} \left(\frac{h_s}{r}\right)^2 - \frac{1}{r-1}.$$

Beweis. Das folgt aus dem Zwischenschritt

$$\varphi(a) = \frac{\sum_{s \in \Sigma} h_s^2 - r}{r \cdot (r-1)}.$$

♦

Korollar 2 (Invarianzsatz 2) *Der Koinzidenzindex eines Textes ist invariant unter monoalphabetischer Substitution und unter Transposition.*

Beweis. Die Anzahl der Zwillingspaare ändert sich nicht. ♦

Im Beispiel des (deutschen) Rheingedichts war $\varphi(a) \approx 0.0715$, beim (englischen) »If ...« $\varphi(c) \approx$

0.0623. Diese Werte entsprechen der Erwartung, da sie ja als Mittelwerte der Autokoinzidenzindizes selbst den typischen Wert der Zeichenkoinzidenz zweier Texte der gleichen Sprache annehmen sollten.

Im [Beispiel](#) des analysierten Geheimtextes war $\varphi(c) \approx 0.0440$. Auch dies ist ein typischer Wert, den wir bald verstehen werden.

Autor: Klaus Pommerening, 24. Februar 2000; letzte Änderung: 25. Mai 2002.
E-Mail an Pommerening »AT« imbei.uni-mainz.de.