Het meten van lexicale variatie

Variatie in het dialectlandschap wordt bepaald door diversiteit (*TTR*), heterogeniteit (*SID*) en entropie (*ETP*). De variatieindex *VAR* wordt nu als volgt berekend:

$$VAR = TTR * SID * ETP$$

a. Diversiteit

A	B	C	D
E	F	G	H
I	J	K	L
M	N	O	P

heeft meer variatie dan

A	A	A	A
A	A	A	A
A	A	A	A
A	A	A	A

Diversiteit is gelijk aan het aantal verschillende termen voor een concept gedeeld door het totale aantal termen in de data set. Het totale aantal termen is gelijk aan het aantal dialectlocaties dat geen 'missing value' heeft. We bereken dus de type/token ratio:

TTR = aantal types / aantal tokens

In het linker plaatje zijn 16 types en 16 tokens, *TTR* is dus gelijk aan 1. In het rechter plaatje is er één type en zijn er 16 tokens. *TTR* is gelijk aan 0.0625. *TTR* variëert dus van maximaal 1 naar minimaal 1/aantal tokens. Een hoge *TTR* duidt op een grote diversiteit.

b. Heterogeniteit

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

heeft meer variatie dan

A	A	B	B
A	A	B	B
C	C	D	D
C	C	D	D

Heterogeniteit meten we met de silhouet index (Rousseeuw, 1987). In ons voorbeeld zijn er vier groepen: de A's, de B's, de C's en de D's. Als de dialectlocaties waar dezelfde term voor het concept gebruikt wordt, een aaneengesloten gebied vormt, zullen de onderlinge geografische afstanden klein zijn (kleine intracluster afstand), maar de afstanden ten opzichte van dialectlocaties waar een andere term gebruikt wordt, groot zijn (intercluster afstand). De silhouet index is een gecombineerde meting van de intracluster

afstand en de intercluster afstand. Voor iedere dialectlocatie *i* waar men term *j* gebruikt is de breedte van het silhouet gelijk aan:

$$s_i^j = \frac{b_i^j - a_i^j}{\max\left\{a_i^j, b_i^j\right\}}$$

Daarbij is a_i^j de gemiddelde geografische afstand tot alle andere dialectlocaties waar term *j* gebruikt wordt. b_i^j wordt als volgt berekend. Voor elke groep – behalve de groep waarin dialectlocatie *i* zich bevindt - wordt berekend de gemiddelde geografische afstand tussen de dialectlocaties in de groep en dialectlocatie *i*. b_i^j is nu de de afstand van de groep met het kleinste groepsgemiddelde ten opzichte van dialectlocatie *i*. b_i^j is dus de geografische afstand tussen dialectlocatie *i* en de groep (waar niet term *j* gebruikt wordt) die geografische het dichtstbijzijnd is.

s_i^j kan variëren van -1 tot +1. Een waarde dicht bij 1 betekent dat dialectlocatie *i* geografisch vooral omringd is door plaatsen waar dezelfde term gebruikt wordt. Is s_i^j gelijk aan 0, dan heeft dialectlocatie *i* voor een deel buren waar dezelfde term gebruikt wordt, en voor een deel buren waar een andere term gebruikt wordt. Als de waarde van s_i^j gelijk is aan -1, dan is dialectlocatie *i* voornamelijk omringd door plaatsen waar niet dezelfde term gesproken wordt.

Het silhouet voor groep *j* wordt berekend door het gemiddelde te nemen van de silhouetten van de dialectlocaties in die groep:

$$S_j = \frac{1}{m_j} \sum_{i=1}^{m_j} s_i^j$$

Ten slotte wordt de globale silhouet index berekend door het gemiddelde van de groepssilhouetten:

$$S = \frac{1}{K} \sum_{j=1}^{K} S_j$$

waarbij K het aantal groepen is. In ons voorbeeld is *K* gelijk aan 4. Omdat elke s_i^j varieert tussen -1 en 1, variëren ook S_j en *S* tussen -1 en 1. We normaliseren *S* naar waarden tussen 0 en 1 als volgt:

SID = -1*((S-1)/2)

Als *SID* gelijk is aan 0, is de gemiddelde variatie rondom de dialectlocaties minimaal, en als *SID* gelijk is aan 1, is de gemiddelde variatie rondom de dialectlocaties maximaal.

c. Entropie

A	A	B	B
A	A	B	B
C	C	D	D
C	C	D	D

heeft meer variatie dan

A	A	A	A
A	A	A	B
A	A	A	B
A	C	C	D

In dit voorbeeld heeft het concept vier verschillende termen (A, B, C en D). In het linker plaatje komt elke term even vaak voor, namelijk vier keer, maar in het rechter plaatje is term A dominant en komt 11 keer voor.

Entropie is op het "fundamenteelste niveau een maat voor de wanorde of de ontaarding in een systeem, of liever de waarschijnlijkheid, als het aantal mogelijke moleculaire configuraties van een macroscopische toestand (in termen van macroscopische grootheden druk, temperatuur, etc.) gedeeld door het totale aantal mogelijke moleculaire configuraties." (definitie Wikipedia: Entropie))

"Entropie is de maat voor informatiedichtheid in een reeks gebeurtenissen. Informatie ontstaat als een gebeurtenis plaatsvindt waarvan vooraf onzeker was of deze daadwerkelijk zou gebeuren." (definitie Wikipedia: Entropie (informatietheorie))

In het linker plaatje is de entropie maximaal, omdat elke term even vaak voorkomt, in het rechter plaatje is de entropie lager. De Entropie *H* wordt gemeten als:

$$H = -\sum_{i=1}^{n} p_i \cdot^2 log(p_i)$$

De entropie voor het linker voorbeeld is 2.000, en voor het rechter voorbeeld is dit: 1.372.

"In het algemeen is het bewijsbaar dat uitvoeren van een experiment dat N mogelijke uitkomsten heeft nooit meer dan een gemiddelde informatie van 2log(N) bit kan opleveren. En deze waarde voor de gemiddelde informatie wordt bereikt als elke uitkomst een even grote kans 1/N heeft." (Wikipedia: Entropie (informatietheorie)). We berekenen nu genormaliseerde entropiewaarden als volgt:

ETP = H / 2log(N)

ETP variëert tussen 0 en 1. De genormaliseerde entropie voor het linker voorbeeld wordt dan gelijk aan 1, en die in het rechter voorbeeld wordt gelijk aan 0.686.