Kwantificering in taal en crypto-analyse [1])

door M. de Vries

Summary

Quantification in language and crypto-analysis

In this paper the statistical properties of different languages are considered and it is shown how every language may be charactarised by a parameter, the so called ,,invariant parameter''. This parameter, and the character of a language in general, can be changed by crypto-transformations. Statistical methods for finding the key of crypto-transformations are discussed and illustrated by examples. It is shown how these methods can be made ineffectual by using special crypto-transformations, eradicating those statistical properties of the language, which constitute systematic deviations from randomness in the numbers and combinations of letters used.

1.1 Inleiding

Het ontcijferen van geheimschriften behoort ongetwijfeld tot de oudste toepassingen van de statistiek. Bijna 500 jaar geleden, in 1466, publiceerde de Italiaanse architect L. B. A l b e r t i het eerste moderne boek over geheimschriften: ,,Modus Scribendi in Ziferas'', waarin hij allerlei bijzonderheden over de statistische structuur van de taal mededeelde.

De pas beginnende en nog experimenterende crypto-analyst van die dagen ontdekte een merkwaardige regelmaat in de schrijftaal. Hij nam waar dat in grote aantallen steekproeven de frequentie-quotienten van de letters in een bepaalde taal nagenoeg constant bleven. Men concludeerde hieruit dat deze statistische regelmaat een karakteriserende eigenschap was van de taal.

Overgebracht in modern statistisch jargon luidt deze conclusie: Er bestaat een bepaalde, onbekende, waarschijnlijkheid waarmee de letters in de taal voorkomen; de frequentie-quotienten worden opgevat als schattingen van de onbekende parameters.

Voor de bescheiden behoeften van die dagen was de kennis van deze frequentie-quotienten of waarschijnlijkheden ruim voldoende om bruikbare resultaten op te leveren. Het lukte heel aardig met zeer elementaire statistische methoden de geheime correspondentie van vriend en vijand te ontcijferen. Toen het echter bekend werd dat de geheimen niet langer geheim waren, bedacht men betere methoden en de crypto-analyst stond voor nieuwe problemen. Men raadpleegde nogmaals de frequentietabellen en ontdekte dat er

---

[1]) Rapport SP 32 van de Statistische Afdeling van het Mathematisch Centrum, Amsterdam. De afdeling staat onder leiding van Prof. Dr D. v a n D a n t z i g.

nog meer taalkenmerken bestonden. Niet slechts de enkele letters van de taal, maar ook de combinaties van twee en drie letters (bigrammen en trigrammen) kwamen voor in bijna vaste verhoudingen. En passant had men nog even de frequenties van begin- en eindletters van de woorden genoteerd, verder nog opgemerkt dat bepaalde uitgangen vaak voorkwamen; en bovendien de verdeling van klinkers en medeklinkers over de woorden onderzocht.

Nu is het natuurlijk niet nodig geheimschriften te ontcijferen als men iets wil weten over de statistische eigenschappen van de taal. Ook zuiver belangstelling voor de taal zelf, als middel tot communicatie en expressie, kan leiden tot een statistische taalanalyse. Literaire belangstelling was de oorzaak van statistische stijlanalyses die een nieuw licht wierpen op de schrijfgewoonten van verschillende auteurs.

Maar historisch ligt de zaak zo dat politieke noodzaak dwong tot het doordringen in de geheimen van mogelijke tegenstanders. Het ontcijferen van vijandelijke geheimschriften werd soms een voorwaarde om te kunnen voortbestaan. Dit was de oorzaak van een intensief taalonderzoek dat tot in deze tijd voortduurt.

2.1 Cryptologie

De cryptologie is een toegepaste wetenschap waarin taal en wiskunde elkaar ontmoeten. Het voornaamste probleem is van semantische aard: *Informatie vervat in geschreven berichten moet verborgen worden.* Om dit vraagstuk geschikt te maken voor een mathematische beschrijving ziet men (voorlopig) af van de semantische aspecten. Een bericht wordt opgevat als een verzameling van verschillende elementen zonder enige betekenis.

Men stelt zich voor dat een imaginaire taalbron berichten letter voor letter voortbrengt volgens een bepaalde waarschijnlijkheidsverdeling, zodanig dat het optreden van een letter afhankelijk is van de voorgaande letters. Een bericht kunnen we dus opvatten als een steekproef uit een stochastisch proces of een tijdreeks.

Als N het aantal letters is dan stellen we een bericht voor als volgt:

$$t_1, t_2, t_3, \ldots, t_N. \tag{1}$$

De waarden die de stochastische variabelen $\underline{t}_i$ kunnen aannemen behoren tot de eindige verzameling:

$$A = (a_1, a_2, \ldots, a_n) \tag{2}$$

De verzameling A heet een normaal alphabet. De elementen a_i zijn de letters van de taal. Iedere permutatie van A:

(3) $$A_i = (a_{i1}, a_{i2}, \ldots, a_{in})$$

wordt een alphabet genoemd.

Het stochastisch proces wordt beschreven door de volgende waarschijnlijkheden:

(a) $p(i)$ is de waarschijnlijkheid dat een letter a_i voorkomt in een bericht.
(b) $p_i(j)$ is de voorwaardelijke waarschijnlijkheid dat een letter a_j volgt op een letter a_i.
De waarschijnlijkheden $p_i(j)$ heten overgangswaarschijnlijkheden. Uit (a) en (b) kunnen we afleiden:
(c) $p(ij) = p(i)\, p_i(j)$ is de waarschijnlijkheid dat een groep van twee letters (bigram) $a_i a_j$ voorkomt in een bericht.

Een zeer belangrijk statistisch kenmerk heeft betrekking op deelverzamelingen van letters uit een bericht. Voor iedere deelverzameling van uitgebreidheid N' die ontstaat door onafhankelijke trekkingen van letters uit een bericht geldt:

(d) de frequentiequotienten $\frac{n_i'}{N'}$ van de letters a'_i naderen tot de waarschijnlijkheden $p(i)$ van de taal

(4) $$\lim \frac{n_i'}{N'} = p(i).$$

2.2 Crypto-transformaties

Onder een crypto-transformatie verstaan we een transformatie die een bericht omzet in een reeks van letters die geen semantische correlatie vertoont met het oorspronkelijke bericht.

Men gaat nu uit van een streng parallelisme tussen de semantische en statistische kenmerken van de taal. Een transformatie van de statistische kenmerken zal een transformatie van de semantische eigenschappen en dus een versluiering van de informatie ten gevolge hebben.

Een crypto-transformatie stellen we voor door T_s. De parameter s van de specifieke transformatie heet sleutel. Doorloopt s de verzameling S van alle mogelijke waarden die s kan aannemen, dan ontstaat een cryptografisch systeem T. Een cryptografisch systeem is dus een verzameling van crypto-transformaties.

Na het toepassen van een crypto-transformatie op een bericht ontstaat een cryptogram:

(5) $$T_s(t_1, t_2, \ldots, t_N) = c_1, c_2, \ldots, c_M.$$

Het toepassen van een crypto-transformatie noemt men gewoonlijk vercijferen.

3.1 Statistische Taalanalyse

In fig. 1 vinden we een tabel en een grafiek van de frequenties van de letters geteld in een willekeurige steekproef van 10 000 letters (gereduceerd tot 1000). De frequentie-quotienten berekend uit deze steekproef zullen we in het vervolg van dit artikel gebruiken als de waarschijnlijkheden waarmee de letters in de taal voorkomen.

	A	B	C	D	E	F	G	H	I	J	K	L	M
n_i:	65	17	13	61	200	6	37	30	66	17	23	35	22
	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
n_i:	108	57	13	0	63	35	58	18	24	18	0	0	14

$n_i = 1000\, p_i$.

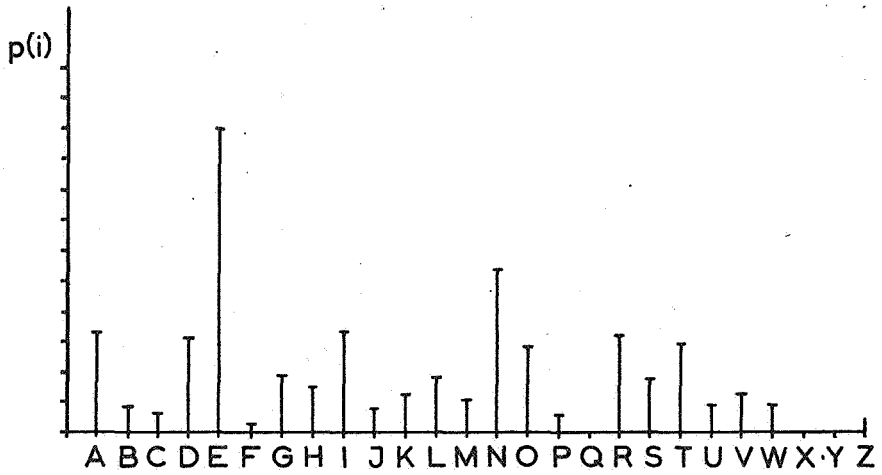

Fig. 1.

Gelukkig zijn letters gemakkelijk bereikbaar materiaal. Iedereen is in staat om zich met behulp van zijn dagblad te voorzien van een steekproef van willekeurige lengte.

Voor een aantal van deze steekproeven kunnen we toetsen of ze uit dezelfde

verdeling afkomstig zijn, bv. met de χ^2-*toets* voor twee steekproeven [1]).

We kunnen namelijk de vraag stellen of de letterfrequenties beïnvloed worden door het onderwerp waaruit de tekst bestaat.

Om deze vraag te beantwoorden werd een aantal malen de χ^2-toets toegepast op een aantal steekproeven. O.a. werd een stukje tekst van V o n d e l vergeleken met een modern krantenbericht. De kleinste overschrijdingskans die werd waargenomen was ongeveer $0{,}01$. Werden daarentegen twee verschillende talen vergeleken dan was de overschrijdingskans 10^{-5} maal kleiner dan bij het vergelijken van steekproeven uit dezelfde taal.

Voor het vergelijken van groepstalen en vaktalen is deze toetsingsmethode niet goed geschikt. Verschillen die voortkomen uit steekproef-onnauwkeurigheden geven al een zeer kleine overschrijdingskans.

Een andere methode om talen te vergelijken, die trouwens meer in overeenstemming is met de verlangens van de crypto-analyst, is met behulp van rangcorrelatie.

De 26 verschillende letters worden in beide steekproeven gerangschikt naar afdalende frequentie. Letters die in frequentie weinig verschillen krijgen hetzelfde rangnummer. De steekproeven worden dan getoetst op overeenkomst.

3.2 **Invariante parameter**

Een groot aantal numerieke gegevens die een verzameling bepalen kunnen we samenvatten in een enkel getal. Meestal worden dan bepaalde eigenschappen van de verzameling verwaarloosd door de nieuwe parameter, maar tevens wordt de nadruk gelegd op andere kenmerken die niet onmiddellijk blijken uit de afzonderlijke getallen.

Voor verschillende toepassingen in de crypto-analyse is een handige taalparameter de som der kwadraten der waarschijnlijkheden waarmee de letters in de taal voorkomen:

$$\kappa_p = \sum_{1}^{26} i\, p_i^2 \tag{6}$$

Voor het Nederlands neemt deze parameter de waarde aan $\kappa = 0{,}0827$.

Deze parameter kunnen we meetkundig interpreteren.

We stellen de taal voor als een punt P in een 26-dimensionale ruimte met

---

[1]) $\chi^2 = M N \sum_{1}^{r} i \frac{1}{m_i + n_i} \left(\frac{m_i}{M} - \frac{n_i}{N}\right)$ waarin:

M = aantal letters in de eerste steekproef
N = aantal letters in de tweede steekproef
m_i = frequentie van i-de letter in de eerste steekproef
n_i = frequentie van i-de letter in de tweede steekproef.
(χ^2 is verdeeld met r-1 vrijheidsgraden).

de kansen p_i als coördinaten. Dan is $\overset{26}{\underset{1}{\Sigma}} p_i^2$ het kwadraat van de afstand van de oorsprong tot P. Het punt ligt verder in het hypervlak: $\overset{26}{\underset{1}{\Sigma}} p_i = 1$.

Uit de definitie volgt dat deze parameter niet eenduidig bepaald is. Alle punten die liggen op de doorsnede van de 25-dimensionale sfeer, met $r^2 = \overset{26}{\underset{1}{\Sigma}} p_i^2$ (r = straal) en het vlak $\overset{26}{\underset{1}{\Sigma}} p_i = 1$, hebben dezelfde afstand tot de oorsprong.

Al deze punten ontstaan ook als we de gegeven waarschijnlijkheden zouden toekennen aan andere letters. We kunnen dit nog anders formuleren: Als we het normale alphabet afbeelden op een permutatie van zichzelf, dan ontstaat een alphabet waarin de gegeven waarschijnlijkheden zijn toegekend aan andere letters.

Het gebruik van de parameter $\varkappa_p$ in de crypto-analyse berust op het feit dat een groot aantal methoden om geheim te schrijven gebaseerd zijn op permutaties van het alphabet en $\varkappa_p$ is invariant voor permutaties van het alphabet.

Een zuivere schatting van $\varkappa_p$ wordt gegeven door de statistische grootheid:

$$K = \Sigma \frac{n_i\,(n_i - 1)}{N\,(N-1)}, \quad \text{waarin} \tag{7}$$

N = aantal letters in de steekproef
n_i = frequentie van de i-de letter.

$$\mathcal{E}\,(\underline{K}) = \varkappa_p \tag{8}$$

De variantie van $\underline{K}$ is

$$\sigma_\kappa^2 = \frac{0{,}0168\,N + 0{,}1182}{N\,(N-1)} \tag{9}$$

Verdere momenten [1]) kunnen eenvoudig berekend worden met behulp van de multinomiale-verdeling. De exacte verdeling is niet eenvoudig te berekenen. Een benadering via de normale verdeling is daarentegen wel af te leiden.

Voor $p_i = \frac{1}{n}$ wordt de verdeling gelijk aan de niet-centrale χ^2-verdeling. In alle andere gevallen hangt de verdeling af van de parameters p_i.

4.1 Crypto-analyse

In een cryptogram is de frequentie verdeling van de letters geheel anders

[1]) Het gereduceerde derde moment is gegeven door:

$$\tilde{\mu}_3 = \frac{(N-2)\,(N-3)\,.\,0{,}1984 + (N-2)\,.\,0{,}3896 + 0{,}0737}{N\,(N-1)}$$

dan in een normale tekst. In een ideaal geheimschrift zijn alle statistische kenmerken verdwenen.

Een eerste benadering hiervan is dat in het cryptogram alle frequentiequotienten gelijk zijn aan $\frac{1}{26}$. De verdeling is rechthoekig.

Het eerste werk van de crypto-analyst is een onderzoek van het onderschepte cryptogram onder de hypothese dat het vercijferd is volgens een ideale methode. Of wat hetzelfde is, onder de hypothese dat het cryptogram ontstaan is door onafhankelijke trekkingen uit een multinomiale-verdeling met 26 kenmerken, alle met dezelfde waarschijnlijkheid.

Het lettermateriaal wordt nu doorzocht of er afwijkingen voorkomen die niet toevallig zijn ontstaan.

Laten we aannemen dat we in het bezit zijn geraakt van 100 letters cryptotekst. Onder de nulhypothese is de verwachting voor een willekeurige letter *3,85*. Sommige letters komen misschien niet voor, andere 6 of 7 maal. Omdat we te maken hebben met een kleine kans is de Poisson-verdeling [1]) een behoorlijke benadering. Met behulp van deze verdeling kunnen we bv. de kans bepalen dat een vooraf bepaalde letter niet voor zal komen.

Deze kans is *0,02*. De verwachting van het aantal letters dat niet voor zal komen in het cryptogram is *0,52*. Wijkt het waargenomen aantal letters dat niet voorkomt significant af van de verwachting dan verwerpt men de hypothese dat het systeem ideaal is. De afwezigheid van bepaalde letters kan een aanwijzing geven over de aard van het toegepaste cryptografisch systeem.

Dezelfde vragen kunnen beantwoord worden voor polygrammen van willekeurige lengte.

4.2 Coincidentietoets-A

De volgende stap die gedaan wordt om een cryptogram te ontcijferen is de analyse van herhalingen.

In iedere taal komen herhalingen voor, zowel van woorden als van lettergroepen. Men wil graag weten of van deze herhalingen in het cryptogram nog iets is overgebleven. Ook in een cryptogram zullen herhalingen voorkomen, onafhankelijk van de herhalingen in het oorspronkelijke bericht, al was het alleen maar omdat er met 26 letters slechts een eindig aantal combinaties van gegeven lengte te construeren zijn. Men kan echter onderzoeken of deze herhalingen bij toeval zijn ontstaan of niet.

We vormen uit een cryptogram alle combinaties van 2 letters. Dit zijn er $\binom{N}{2}$, als N weer het aantal letters in het cryptogram voorstelt. Bestaat een

[1]) E. C. Molina: Poisson's Experimental Binomial Limit, Tabel I.

dergelijke combinatie uit 2 gelijke letters, dan spreekt men van een *coincidentie*. Als het geheimschrift ideaal is, dan is de kans dat in één trekking uit het universum twee gelijke, maar overigens willekeurige, letters worden getrokken, gelijk aan de som van de individuele kansen:

$$\varkappa_r = \sum_1^{26} \left(\frac{1}{26^2}\right) = 0{,}0385 \tag{10}$$

Dit is de kans op een coincidentie. De grootheid $\varkappa_r$ heet *de coincidentieconstante voor rechthoekig verdeelde tekst*.

De verwachting van het aantal coincidenties is:

$$\varkappa_r \frac{N(N-1)}{2}. \tag{11}$$

Het waargenomen aantal is:

$$\sum_1^{26} \frac{n_i(n_i-1)}{2} \tag{12}$$

als n_i de frequentie van de letter a_i voorstelt.

Met behulp van de Poisson-verdeling worden de waarschijnlijkheden bepaald van het voorkomen van 0, 1, 2, ... coincidenties.

Voor normale tekst is de kans op een coincidentie:

$$\varkappa_p = \Sigma p_i^2 = 0{,}0827 \tag{13}$$

De grootheid $\varkappa_p$ heet de *coincidentieconstante voor normale tekst*.

Hier zien we de invariante parameter Σp_i^2, die we op theoretische gronden hebben ingevoerd, terugkeren als de kans op een coincidentie. In de volgende paragrafen zullen we deze parameter nog enige malen ontmoeten in toetsingsmethoden die alle dezelfde theoretische achtergrond bezitten.

Er is geen reden de coincidentietoets te beperken tot enkele letters. Als m de lengte is van het polygram, dan geldt voor rechthoekig verdeelde cryptotekst:

$$\varkappa_{rm} = \binom{N-m+1}{2} \sum_1^{26^m} \left(\frac{1}{26^m}\right)^2. \tag{14}$$

4.3 Coincidentietoets-B

Een van de belangrijkste statistische technieken in de crypto-analyse staat bekend als coincidentietoets-B.

Deze toetsingsmethode heeft als theoretische achtergrond de volgende overwegingen:

a) Als twee voldoende lange reeksen van letters worden gesuperponeerd, dan zullen in een aantal gevallen gelijke letters uit de bovenste en onderste rij boven elkaar staan.
   Dergelijke paren gelijke letters noemen we weer coincidenties..
b) Bestaan beide reeksen van letters uit rechthoekig verdeelde tekst, dan is de kans op een coincidentie: $\varkappa_r = 0{,}0385$.
c) Bestaan beide reeksen van letters uit normale tekst, dan is de kans op een coincidentie: $\varkappa_p = 0{,}0827$.

Om de toepassing van deze methode toe te lichten zullen we een voorbeeld geven van een crypto-transformatie waarbij deze toets meestal tot snelle resultaten leidt.

We stellen ons een crypto-transformatie voor T_{si} (de sleutel s blijft constant), die afhangt van een ,,tijd''-parameter i ($i = 1, \ldots, N$).

De transformatie bestaat dan uit een reeks van transformaties, die worden toegepast op de afzonderlijke letters:

$$T_{s1}, T_{s2}, \ldots, T_{sN}. \tag{15}$$

Als we de tekst voorstellen door:

$$t_1, t_2, \ldots, t_N. \tag{16}$$

dan is het cryptogram te schrijven als:

$$T_{s1}\, t_1, T_{s2}\, t_2, \ldots, T_{sN}\, t_N. \tag{17}$$

Een andere normale tekst van dezelfde lengte:

$$u_1, u_2, \ldots, u_N, \tag{18}$$

wordt met dezelfde transformatie vercijferd:

$$T_{s1}\, u_1, T_{s2}\, u_2, \ldots, T_{sN}\, u_N. \tag{19}$$

Geldt nu:

$$t_k = u_k$$

dan volgt:

$$T_{sk}\, t_k = T_{sk}\, u_k.$$

Superponeren wij de twee cryptogrammen zo dat de transformaties in de bovenste rij corresponderen met de transformaties in de onderste rij, dan is het resultaat gelijk aan een superpositie van twee normale teksten. De verwachting van het aantal coincidenties is: $\varkappa_p N$.

Een onjuiste superpositie van twee cryptogrammen vercijferd met dezelfde sleutel zal ook aanleiding geven tot het optreden van coincidenties; maar deze

zijn dan geheel van toevallige aard. Het aantal zal dan niet ver liggen van $\varkappa_r N$.

In de crypto-analytische practijk komt het superponeren herhaaldelijk voor. Allereerst als methode om te weten te komen of meerdere cryptogrammen met dezelfde sleutel zijn vercijferd. Een ander voorbeeld dat terug te brengen is tot het vorige is de ontcijfering van een cryptogram vercijferd met een sleutel met lange periode. Dit betekent dat het aantal transformaties T_{si} korter is dan de tekst. Heeft men enige aanwijzingen omtrent de lengte van de sleutel en voldoende materiaal, dan verdeelt men de tekst in stukken en probeert de juiste superpositie te vinden. Met andere, niet altijd statistische, methoden worden de stukken dan simultaan ontcijferd.

4.4 Toets voor monoalphabeticiteit

Het is zeer vaak van belang te weten of een letterverdeling afkomstig is uit een alphabet. Onder een alphabet verstaan we een willekeurige permutatie A_i van A.

Bij grote aantallen letters is het vergelijken van de frequentiequotienten met de waarschijnlijkheid $p(i)$ meestal voldoende.

Voor kleine aantallen waarnemingen is vergelijken niet meer voldoende om tot een conclusie te komen. Nog sterker geldt dit als de letters een permutatie vormen van het normale alphabet. De toetsingsmethode die men gebruikt om op deze vraag een antwoord te geven is een toepassing van de in paragraaf 3.2 ontwikkelde theorie. Daar is afgeleid dat $\varkappa$ invariant is voor permutaties van het alphabet.

De verwachting van $\underline{K}$ is gelijk aan $\varkappa_p$.

Als K niet te veel afwijkt van $\varkappa_p$ dan mogen we aannemen dat de steekproef afkomstig is uit een alphabet.

4.5 H-toets

Aan de toepassingen van $\sum_1^{26} p_i^2$ kunnen we er nog een paar toevoegen.

In een ideaal geheimschrift is $\sum_1^{26} p_i^2 = 0{,}0385$, en in een normale tekst: $\sum_1^{26} p_i^2 = 0{,}0827$.

In een cryptogram zal de waarde van K tussen deze grenzen liggen. $\sum_1^{26} p_i^2$ is dus ook een *maat* voor de *doeltreffendheid* van het gebruikte geheimschrift.

Meestal gebruikt men in plaats van K de grootheid $H = 26K$ als maat voor de doeltreffendheid van een cryptografisch systeem.

Voor een rechthoekig verdeelde tekst is:

(20) $$H = 26\,\varkappa_r = 1$$

Voor normale tekst (Nederlands):

(21) $$H = 26\,\varkappa_p = 2{,}15$$

Een belangrijke toepassing van de *H-toets* is de volgende:

Een tekst wordt vercijferd in stappen door het toepassen van verschillende crypto-transformaties na elkaar. Bij de ontcijfering probeert men deze transformaties achtereenvolgens te elimineren. Men hoopt nu dat met een eliminatie van een transformatie een stijging van H gepaard zal gaan. Meestal echter is de toepassing van de toets zeer moeilijk omdat de steekproeffouten vaak groter zijn dan de variatie van H.

4.6 Twee steekproeventoets

In vele gevallen weten we van twee verdelingen dat het steekproeven zijn uit een normaal alphabet of een permutatie daarvan. Het is dan van belang te weten of beide alphabetten dezelfde permutatie van A voorstellen of niet. Bij grote aantallen letters is het vergelijken van de frequentie-verdelingen voldoende. Voor kleine aantallen waarnemingen heeft men verschillende methoden bedacht die aanwijzen of twee steekproeven afkomstig zijn uit dezelfde verdeling.

Volgens paragraaf 4.5 mogen we schrijven:

(22) $$\Sigma\,(n_i^2 - n_i) \approx \varkappa_p N^2 - \varkappa_p N;$$

(23) $$\Sigma\, n_i = N;$$

(24) $$\Sigma\, n_i^2 \approx \varkappa_p N^2 - \varkappa_p N + N$$

We veronderstellen nu dat twee monoalphabetische verdelingen tot hetzelfde alphabet behoren. Worden de twee verdelingen gecombineerd, dan moet, onder de hypothese van gelijkheid, ook de nieuwe verdeling tot datzelfde alphabet behoren.

De letters van de eerste verdeling worden aangeduid met a_{i1} en hun frequenties met n_{i1}; voor de tweede verdeling resp. a_{i2} en n_{i2}. Het aantal letters in de eerste verdeling is N_1 en in de tweede verdeling N_2. Dan geldt:

(25) $$\Sigma\,(n_{i1}^2 + n_{i2})^2 \approx \varkappa_p\,(N_1 + N_2)^2 - \varkappa_p\,(N_1 + N_2) + (N_1 + N_2)$$

(26) $$\Sigma n_{i1}^2 + \Sigma n_{i2}^2 + 2\,\Sigma n_{i1}\, n_{i2} \approx \varkappa_p\,(N_1^2 + N_2^2 + 2N_1N_2) - \varkappa_p\,(N_1 + N_2) + (N_1 + N_2)$$

uit (24) volgt:

(27) $\Sigma n_{i1}^2 \approx \varkappa_p N_1^2 - \varkappa_p N_1 + N_1$

(28) $\Sigma n_{i2}^2 \approx \varkappa_p N_2^2 - \varkappa_p N_2 + N_2$

Na aftrekken links en rechts houden we over:

(29) $2 \Sigma n_{i1} n_{i2} \approx \varkappa_p (2N_1 N_2)$

(30) $\chi = \dfrac{\Sigma n_{i1} n_{i2}}{N_1 N_2} \approx \varkappa_p = 0{,}0827$

De statistische grootheid χ stelt ons in staat te zien of twee steekproeven afkomstig zijn uit hetzelfde alphabet.

Men noemt deze toets gewoonlijk de *χ-toets.*

5.1 Toepassingen

In deze paragraaf zullen enkele van de beschreven statistische methoden toegepast worden bij de ontcijfering van een cryptogram. We geven eerst een beschrijving van het gebruikte cryptografisch systeem.

De letters van het alphabet worden opgevat als de elementen van een eindige groep. Een samenstellingsvoorschrift is gegeven dat we optelling noemen. Hetzelfde resultaat bereiken we door de letters (in een bepaalde volgorde) één-éénduidig af te beelden op de natuurlijke getallen 0 — 25. De letters van het alphabet stellen dan de restklassen mod 26 voor. In het voorbeeld zijn de letters in hun normale volgorde afgebeeld op de getallen 0—25.

Om een goede vercijfering tot stand te brengen met dit systeem moet de sleutel zeer lang zijn. Dit is o.a. te bereiken door het bericht trapsgewijze te vercijferen met betrekkelijk korte sleutels. De lengte van de resulterende sleutel is dan hun kleinste gemene veelvoud. Het tot stand brengen van een dergelijke vercijfering met behulp van potlood en papier is een zeer tijdrovende bezigheid die bovendien grote kans biedt op het maken van fouten. Verschillende typen apparaten waarmee berichten mechanisch of electrisch vercijferd worden berusten op het beginsel van herhaalde toepassing van niet te lange sleutels.

De grote hoeveelheid tekst die nodig is om resultaten te verkrijgen verhindert een demonstratie van de statistische ontcijfering van een machinaal vercijferd cryptogram.

5.2 Voorbeeld

Het volgende cryptogram is onderschept en wordt onderworpen aan een statistische analyse:

	(10)		(20)		(30)		(40)		(50)
FATSM	EEXVR	WPZWK	MXVKX	TMTWE	FLXTO	WMGOP	RXHRF	ILBMS	FQRML
	(60)		(70)		(80)		(90)		(100)
FDEGO	WPKSG	KMBLV	XTZKZ	EXUWK	SMMWO	CEFPI	ETWDM	YBUGO	VSDTZ
	(110)		(120)		(130)		(140)		(150)
MTTHG	ZMELG	EUWPO	ZDXGI	DEXSK	GUZSG	SRUTH	TIXHE	KASZE	GZWIL
	(160)		(170)		(180)		(190)		(200)
DETTB	OWLBE	DUSMP	WOHXS	WPTIJ	DEGMF	DCRXQ	NOVSD	GXALG	SHOHR
	(210)		(220)		(230)		(240)		(250)
WMFWW	SBQEG	ZXACA	IKVPW	UAPKW	UXAGB	KPMVX	DAGAX	QYQSM	IXANT
	(260)		(270)		(280)		(290)		(300)
VDOQR	XLRHO	NLAID	MKDUG	LAZME	KDDMY	EZSYW	JOLVW	GDXHW	KMIUE
	(310)		(320)		(330)		(340)		(350)
EGPRM	NWMJG	KMWZE	XVKXU	EXFIL	JLQOO	MJGLR	CHMLB	NXZXY	ZEGTI
	(360)		(370)		(380)		(390)		(400)
FDXWG	TZTJH	NDMDB	ROOPE	VYCBT	HIFCS	CKXBX	ALNLI	ODDWP	IQIZM
	(410)		(420)		(430)		(440)		(450)
FUJLS	IPXVN	TVHOF	ATSHU	EBVNC	YMSGT	UTACD	IXNIC	JVKPM	WLTHR
	(460)								
KYOKU	SMQIC.								

Het totaal aantal letters in het cryptogram is 460.
De frequentieverdeling van de letters is:

	A	B	C	D	E	F	G	H	I	J	K	L	M
n_i:	15	12	11	23	24	13	24	15	20	8	20	20	31
	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
n_i:	10	20	15	9	13	21	24	15	15	26	30	9	17

	A	B	C	D	E	F	G	H	I	J	K	L	M
	0	1	2	3	4	5	6	7	8	9	10	11	12
	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
	13	14	15	16	17	18	19	20	21	22	23	24	25

Een eindige rij letters voorgesteld door:

$$s_1, s_2, \ldots, s_k \tag{31}$$

is gegeven als sleutel. Deze letters worden opgeteld bij de letters van de tekst op de volgende manier:

$$\begin{array}{l} t_1, t_2, t_3, \ldots, t_k, t_{k+1}, \ldots, t_{2k}, t_{2k+1}, \ldots, t_{3k}, t_{3k+1} \ldots \\ +\ s_1, s_2, s_3, \ldots, s_k, s_1, \ldots, s_k, s_1, \ldots, s_k, s_1 \ldots \\ \hline c_1, c_2, c_3, \ldots, c_k, c_{k+1}, \ldots, c_{2k}, c_{2k+1}, \ldots, c_{3k}, c_{3k+1} \ldots \end{array}$$

Deze transformatie is een voorbeeld van de transformatie beschreven in paragraaf 4.3, (15), (16) en (17). De parameters van de transformatie zijn de letters $s_1, \ldots, s_k$. De operatie is de optelling mod 26.

Om mnemotechnische redenen kiest men voor de eindige sleutel vaak een woord of een zin. Dit is echter niet noodzakelijk; iedere rij letters kan als sleutel dienst doen.

Uit (31) blijkt dat de sleutel cyclisch wordt gebruikt. Het vercijferen van een bericht met het beschreven cryptografisch systeem kunnen we opvatten als een stochastisch proces dat systematisch gestoord wordt [1]).

We mogen deze beschouwingswijze ook omkeren en het cryptogram opvatten als een stochastisch proces of tijdreeks ontstaan uit een cyclische component en een stochastische storing met bekende verdeling. De analyse van een dergelijk cryptogram komt dus neer op het *splitsen van een tijdreeks in een cyclische en een stochastische component.*

Al is deze verdeling veel vlakker dan van een normale tekst, de afwijkingen van de rechthoekige verdeling zijn zo duidelijk dat het niet nodig is deze nog afzonderlijk aan te tonen.

We zullen de coincidentietoets-A gebruiken om te onderzoeken of het aantal coincidenties van tetragrammen significant afwijkt van de verwachting.

In de tekst vinden we de volgende tetragram-coincidenties:

FATS	op de plaatsen	1 en 415
XVKX	,,	17 en 316
OVSD	,,	95 en 187

De verwachting van het aantal tetragram-coincidenties is:

$$(32) \qquad \binom{N-4+1}{2} \Sigma \left(\frac{1}{26^4}\right)^2 = \binom{457}{2} \frac{1}{26^4} = 0{,}23.$$

De kans op het voorkomen van 3 coincidenties in het gegeven cryptogram onder de hypothese dat het ontstaan is door onafhankelijke trekkingen uit een rechthoekige verdeling is *0,0015*.

De gevonden coincidenties zijn dus met vrij grote zekerheid niet toevallig ontstaan.

Deze coincidenties worden nu gebruikt om te onderzoeken of het cryptogram vercijferd is met een periodieke sleutel. Als we inderdaad te maken hebben met een periodieke sleutel, en de letters FATS zijn op beide plaatsen in het cryptogram ontstaan door vercijfering van gelijke letters uit de normale tekst met dezelfde sleutelletters, dan is de afstand tussen beide tetragrammen een veelvoud van de sleutellengte. We vinden:

[1]) Deze interpretatie van het vercijferen is afkomstig uit de informatie-theorie.

FATS: 415–1 = 414 = 2.9.23
XVKX: 316–17 = 299 = 13.23
OVSD: 187–95 = 92 = 4.23

In alle drie gevallen vinden we dat de afstand tussen de coincidenties een veelvoud van 23 is. Voorlopig nemen we aan dat de tekst inderdaad vercijferd is met een sleutel van de lengte 23.

Er moet de nadruk op gelegd worden dat dit een uiterst eenvoudig voorbeeld is. In de meeste gevallen is de afstand tussen de coincidenties een functie van de sleutellengte. Deze functie is uit een klein aantal waarnemingen niet altijd eenduidig te bepalen.

Het cryptogram wordt uitgeschreven in rijen van 23 letters (tabel 1):

TABEL 1

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
F	A	T	S	M	E	E	X	V	R	W	P	Z	W	K	M	X	V	K	X	T	M	T
W	E	F	L	X	T	O	W	M	G	O	P	R	X	H	R	F	I	L	B	M	S	F
Q	R	M	L	F	D	E	G	O	W	P	K	S	G	K	M	B	L	V	X	T	Z	K
Z	E	X	U	W	K	S	M	M	W	O	C	E	F	P	I	E	T	W	D	M	Y	B
U	G	O	V	S	D	T	Z	M	T	T	H	G	Z	M	E	L	G	E	U	W	P	O
Z	D	X	G	I	D	E	X	S	K	G	U	Z	S	G	S	R	U	T	H	T	I	X
H	E	K	A	S	Z	E	G	Z	W	I	L	D	E	T	T	B	O	W	L	B	E	D
U	S	M	P	W	O	H	X	S	W	P	T	I	J	D	E	G	M	F	D	C	R	X
Q	N	O	V	S	D	G	X	A	L	G	S	H	O	H	R	W	M	F	W	W	S	B
Q	E	G	Z	X	A	C	A	I	K	V	P	W	U	A	P	K	W	U	X	A	G	B
K	P	M	V	X	D	A	G	A	X	Q	Y	Q	S	M	I	X	A	N	T	V	D	O
Q	R	X	L	R	H	O	N	L	A	I	D	M	K	D	U	G	L	A	Z	M	E	K
D	D	M	Y	E	Z	S	Y	W	J	O	L	V	W	G	D	X	H	W	K	M	I	U
E	E	G	P	R	M	N	W	M	J	G	K	M	W	Z	E	X	V	K	X	U	E	X
F	I	L	J	L	Q	O	O	M	J	G	L	R	C	H	M	L	B	N	X	Z	X	Y
Z	E	G	T	I	F	D	X	W	G	T	Z	T	J	H	N	D	M	D	B	R	O	O
P	E	V	Y	C	B	T	H	I	F	C	S	C	K	X	B	X	A	L	N	L	I	O
D	D	W	P	I	Q	I	Z	M	F	U	J	L	S	I	P	X	V	N	T	V	H	O
F	A	T	S	H	U	E	B	V	N	C	Y	M	S	G	T	U	T	A	C	D	I	X
N	I	C	J	V	K	P	M	W	L	T	H	R	K	Y	O	K	U	S	M	Q	I	C

Schrijven we een *normale tekst* in rijen van 23 letters dan vormen de letters in de kolommen steekproeven uit normale tekst.

In paragraaf 2.1 (d) is vastgesteld dat voor een dergelijke steekproef geldt dat de frequentie-quotienten $\frac{n_i'}{N'}$ naderen tot de waarschijnlijkheden $p(i)$ van de taal.

Als onze veronderstelling juist is en de tekst is vercijferd met een sleutel

van 23 letters, betekent dit dat in iedere kolom van tabel 1 letters staan die vercijferd zijn met dezelfde (vooralsnog onbekende) sleutelletter.

Uit de algebraische eigenschappen van de transformatie volgt onmiddellijk dat de vercijfering van de letters uit een normaal alphabet een permutatie van dit alphabet tengevolge heeft. De frequentiequotienten veranderen niet van waarde, maar zijn toegevoegd aan andere letters. Men zegt meestal dat de letters „behoren" tot een bepaald alphabet.

De vraag of de letters tot een alphabet behoren toetsen we met de toets voor monoalphabeticiteit. De frequenties van de letters voor ieder van de 23 kolommen van tabel 1, vinden we in tabel 2.

TABEL 2

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
A	–	2	–	1	–	1	1	1	2	1	–	–	–	–	1	–	–	2	2	–	1	–	–
B	–	–	–	–	–	1	–	1	–	–	–	–	–	–	–	1	2	1	–	2	1	–	3
C	–	–	1	–	1	–	1	–	–	–	2	1	1	1	2	–	–	–	–	1	1	–	1
D	2	3	–	–	–	5	1	–	–	–	–	1	1	–	–	1	1	–	1	2	1	1	1
E	1	7	–	–	1	1	5	–	–	–	–	–	1	1	–	3	1	–	1	–	–	3	–
F	3	–	1	–	1	1	–	–	–	2	–	–	–	1	–	–	1	–	2	–	–	–	–
G	–	1	3	1	–	–	1	3	–	2	4	–	1	1	3	–	2	1	–	–	–	1	–
H	1	–	–	–	1	1	1	1	–	–	–	2	1	–	4	–	–	1	–	1	–	1	–
I	–	2	–	–	3	–	–	–	2	–	2	–	1	–	1	2	–	1	–	–	–	5	–
J	–	–	–	2	–	–	–	–	–	3	–	1	–	2	–	–	–	–	–	–	–	–	–
K	1	–	1	–	–	2	–	–	–	2	–	2	–	3	2	–	2	–	2	1	–	–	2
L	–	–	1	3	1	–	–	–	1	2	–	3	1	–	–	–	2	2	2	1	1	–	–
M	–	–	4	–	1	1	–	2	6	–	–	–	3	–	2	3	–	3	–	1	4	1	–
N	1	1	–	–	–	–	1	1	–	1	–	–	–	–	–	1	–	–	3	r	–	–	–
O	–	–	2	–	–	1	3	1	1	–	3	–	–	1	–	1	–	1	–	–	–	1	5
P	1	1	–	3	–	–	1	–	–	–	2	3	–	–	1	2	–	–	–	–	–	1	–
Q	4	–	–	–	–	2	–	–	–	–	1	–	1	–	–	–	–	–	–	–	1	–	–
R	–	2	–	–	2	–	–	–	–	1	–	–	3	–	–	2	1	–	–	–	1	1	–
S	–	1	–	2	3	–	2	–	2	–	–	2	1	4	–	1	–	–	1	–	–	2	–
T	–	–	2	1	–	1	2	–	–	1	3	1	1	–	1	2	–	2	1	2	3	–	1
U	2	–	–	1	–	1	–	–	–	–	1	1	–	1	–	1	1	2	1	1	1	–	1
V	–	–	1	3	1	–	–	–	2	–	1	–	1	–	–	–	–	3	1	–	2	–	–
W	1	–	1	–	2	–	–	2	3	4	1	–	1	3	–	–	1	1	3	1	2	–	–
X	–	–	3	–	3	–	–	5	–	1	–	–	–	1	1	–	6	–	–	6	–	1	4
Y	–	–	–	2	–	–	–	1	–	–	–	2	–	–	1	–	–	–	–	–	–	1	1
Z	3	–	–	1	–	2	–	2	1	–	–	–	2	1	1	–	–	–	–	1	1	1	–

Voor iedere kolom is $N = 20$. Het gemiddelde van K voor alle 23 kolommen is $0{,}073$. We mogen aannemen dat dit gemiddelde bij benadering normaal verdeeld is. De spreiding is $0{,}007$. De afwijking van de verwachting $\mathcal{E}(\underline{K}) = 0{,}0827$ is niet significant; we mogen dus aannemen dat de verdeling in

alle kolommen tot een alphabet behoren en dat de lengte van de sleutel inderdaad gelijk is aan 23.

Hetzelfde resultaat had bereikt kunnen worden door toepassing van coincidentietoets-B. Tabel 1 stelt dan de superposities voor van stukken cryptogram van 23 letters. De coincidenties worden dan geteld tussen alle combinaties 2 aan 2 van de gesuperponeerde rijen. Het resultaat zal gelijk zijn aan de gemiddelde waarde van K.

Nu bestaat de mogelijkheid dat verschillende kolommen vercijferd zijn met dezelfde sleutelletter. Dit betekent dat ze behoren tot hetzelfde alphabet. Als het aantal sleutelletters groter is dan 26 dan weten we zeker dat een aantal kolommen met dezelfde sleutelletter vercijferd zijn. We onderzoeken dit met de *χ-toets.*

TABEL 3

kolommen	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	38	13	20	18	83	25	25	15	33	23	28	38	33	25	25	28	15	48	33	33	25	23
2		8	20	50	60	135	15	25	18	25	20	50	30	15	90	35	18	45	18	18	90	9
3			28	45	28	35	90	80	53	70	20	53	40	65	45	75	68	30	75	70	40	70
4				40	13	28	18	40	40	43	75	40	38	28	25	23	40	38	18	38	28	10
5					10	33	55	68	43	28	40	55	60	35	48	65	35	38	60	35	78	35
6						42	25	28	20	23	33	42	40	35	40	35	40	40	53	48	33	50
7							23	23	15	58	33	33	45	33	65	20	28	38	23	23	65	48
8								57	40	42	10	40	42	65	23	100	40	28	100	40	38	65
9									40	40	18	75	48	42	63	13	90	48	28	95	55	13
10										38	35	40	73	33	13	63	35	75	42	38	10	28
11											40	35	25	57	45	25	53	20	25	42	48	53
12												25	45	50	30	30	30	38	33	23	30	25
13													30	48	60	25	50	25	33	70	53	8
14														33	23	50	18	57	38	23	38	45
15															30	40	45	18	48	28	45	30
16						(χ is met 1000 vermenigvuldigd)										23	48	28	33	57	75	30
17																	28	40	120	23	33	93
18																		45	30	83	28	30
19																			—	—	—	—
20																				—	—	—
21																					—	—
22																						—
23																						

Alle alphabetten worden met elkaar vergeleken. De resultaten zijn te vinden in tabel 3. De onderstreepte getallen zijn significant [1]) en geven aan dat de

[1]) Significant wordt hier gebruikt in een betekenis die afwijkt van het gebruik in de mathematische statistiek. De verdeling van χ is nog nooit berekend. We bedoelen dat onder de hypothese dat de twee steekproeven afkomstig zijn uit hetzelfde alphabet χ niet "*significant*" afwijkt van de verwachting.

twee kolommen hoogstwaarschijnlijk letterverdelingen bezitten uit hetzelfde alphabet. We vinden dat de volgende kolommen dezelfde verdeling bezitten:

a) 1=6 b) 2=7=16=22 c) 3=8=9=17=20 d) 4=12 e) 10=14=19
16=22 8=17=20
5=22 17=20
9=13=18=21
18=21

Na eliminatie van enkele twijfelachtige gevallen houden we over:

1=6	4=12
2=7=16	10=14=19
3=8=17=20	18=21

De vermoedelijk gelijke verdelingen worden gecombineerd. Het resultaat wordt getoond in tabel 4.

TABEL 4

	1 + 6	*2 + 7 + 16*	*3 + 8 + 17 + 20*	*18 + 21*	*4 + 12*	*10 + 14 + 19*
A	1	3	1	3	1	3
B	1	1	5	2	–	–
C	–	1	2	– 1	1	1
D	7	5	3	1	1	1
E	2	15	1	–	–	2
F	4	–	2	–	–	5
G	–	2	8	1	1	3
H	2	1	2	1	2	–
I	–	4	–	1	–	–
J	–	–	–	–	3	5
K	3	–	4	–	2	7
L	–	–	4	3	6	4
M	1	3	7	7	–	–
N	1	3	2	–	–	4
O	1	4	3	1	–	1
P	1	4	–	–	6	–
Q	6	–	–	1	–	–
R	–	4	1	1	–	1
S	–	4	–	–	4	5
T	1	4	4	5	2	3
U	3	1	2	3	2	2
V	–	–	1	5	3	1
W	1	–	3	3	–	10
X	–	–	20	–	3	2
Y	–	–	–	–	2	–
Z	5	–	3	1	–	1

Deze combinaties kunnen voor een deel al herkend worden als cyclische permutaties van het normale alphabet [1]). Om dit vermoeden te bevestigen zouden we de verdelingen kunnen vergelijken met een standaard-alphabet (fig. 1). Hiervoor zouden we weer de χ-*toets* kunnen gebruiken. Het onderzochte alphabet moet dan zo lang cyclisch gepermuteerd worden tot χ een significante waarde aanneemt. In de practijk echter kan men uit enkele karakteristieke eigenschappen van de frequentie-verdeling van het alphabet onmiddellijk zien welk alphabet we moeten nemen. Het interval tussen de hoge frequenties van de E en de N bijv. is bij een cyclische permutatie constant.

We vinden tenslotte:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
M	A	T	H	—	M	A	T	—	S	—	H	—	S	—	A	T	I	S	T	I	—	—

Deze rij letters vullen we onmiddellijk aan tot de sleutelzin:

MATHEMATISCHE STATISTIEK

Achteraf blijkt dat de combinatie van een paar ongelijke frequentieverdelingen de analyse niet heeft verstoord.

De ontcijfering van het cryptogram is nu zeer eenvoudig en wordt aan de lezer overgelaten.

Zonder enige veronderstellingen te maken over de inhoud van de oorspronkelijke tekst zijn we er in geslaagd met behulp van uitsluitend statistische methoden het cryptogram te ontcijferen.

6.1 Statistische cryptografie

Het belangrijkste vraagstuk in de cryptologie betreft de doeltreffendheid van een cryptografisch systeem. Het is toch de bedoeling van een geheimschrift dat het een geheim verbergt. Tot nu toe hebben we uitsluitend aandacht besteed aan de ontcijfering van een cryptogram, alsof we al bij voorbaat aannamen dat het onmogelijk is een onontcijferbaar geheimschrift te construeren. Nu bestaat er een beroemde uitspraak van Edgar Allan Poe die beweerde dat ieder geheimschrift te ontcijferen is; maar Poe was ook geen crypto-analyst.

Het is niet zo moeilijk een systeem te bedenken dat werkelijk onaantastbaar is voor de meest geraffineerde crypto-analytische methoden. Het is daarentegen een bijna onoplosbaar probleem een geheimschrift te construeren, dat

---

[1]) Uit de algebraische eigenschappen van de transformatie blijkt dat het aantal plaatsen dat het alphabet cyclisch verschoven is gelijk is aan de waarde van het getal waarop de sleutelletter is afgebeeld (zie paragraaf 5.1).

doeltreffend, eenvoudig te leren en vlot te hanteren is, en waarbij de kans op het maken van fouten tot een minimum is teruggebracht.

Een ideaal geheimschrift is een systeem waarin o.a. alle statistische kenmerken van de taal verdwenen zijn. Aan deze voorwaarde voldoet nog geen enkel bestaand cryptografisch systeem.

Bij het ontwikkelen van nieuwe systemen zal men rekening moeten houden met de statistische kenmerken van de taal.

In het Nederlands bijvoorbeeld waarin de E het meest voorkomt en een frequentie heeft van *20%* zouden we een reductiefactor willen invoeren die deze frequentie verlaagde tot *3,85%*. Terwijl we de frequentie van de Z, *1,4%*, willen expanderen tot *3,85%*. Als we hierin slagen hebben alle letters dezelfde frequentie en is in eerste benadering een statistisch taalkenmerk verdwenen.

6.2 **Aselecte sleutel**

Toen men aan het eind van de vorige eeuw enig inzicht begon te krijgen in de mathematische achtergronden van de cryptologie, begreep men dat er een absoluut onaantastbaar cryptografisch systeem bestond. Men realiseerde zich echter niet dat de toepassing van de uiterst simpele theorie zou stranden op technische moeilijkheden.

De operationele kant van de methode is gelijk aan de toepassing beschreven in paragraaf 5.1. De sleutel is echter geen eindige rij letters die periodiek wordt toegepast op het bericht, doch een *aselecte rij van letters* minstens net zo lang als het bericht zelf.

Uit de statistische eigenschappen van deze sleutel en de algebraische eigenschappen van het cryptografisch systeem volgt eveneens de aselectheid van het cryptogram. De frequentie-verdeling van de letters is bijna rechthoekig, bovendien bestaat er geen afhankelijkheid meer tussen opvolgende letters.

Vanzelfsprekend mag een dergelijke sleutel na toepassing op een bericht niet nogmaals gebruikt worden voor de vercijfering van een ander bericht. In paragraaf 4.3 is beschreven hoe twee of meer cryptogrammen vercijferd met dezelfde sleutel simultaan ontcijferd kunnen worden. Aan deze voorwaarde opgelegd aan het gebruik dankt het systeem de naam van *„eenmaal gebruikte sleutel"* [1]).

De grote moeilijkheid schuilt echter in het construeren van dergelijke sleutels. Voor toepassingen in de practijk moet men de beschikking hebben over grote aantallen sleutels, of, wat op hetzelfde neerkomt, één zeer lange sleutel waaruit men voor ieder bericht een stuk kiest.

---

[1]) Andere namen voor hetzelfde cryptografisch systeem zijn: eenmaal gebruikt opteltal, eenmaal gebruikte subsitutie, one time pad.

Men kan bijvoorbeeld reeksen van aselecte letters vastleggen in boekjes en daaruit de sleutel kiezen. Met behulp van een eenvoudige indicator verborgen in het cryptogram geeft men zijn correspondent te kennen waar hij de sleutel in het boekje moet zoeken.

De toepassingen van een dergelijke aselecte sleutel zijn echter beperkt. Het vercijferen is een omslachtig en tijdrovend werk dat grote kans biedt op het maken van fouten.

Men heeft dan ook gezocht naar middelen om het vercijferingsproces te mechaniseren. Sedert het begin van deze eeuw heeft een groot aantal constructeurs zich met deze mechanisatie bezig gehouden. De bekendste cryptografen zijn de ,,*Hagelin*'', van Zweedse makelij en de Duitse ,,*Enigma*'', waarvan de eerste mechanisch en de tweede electrisch vercijfert. Beide toestellen construeren een zeer lange sleutel van millioenen letters, zodat een tweemalig gebruik van dezelfde sleutel niet hoeft voor te komen. Helaas ontbreekt in beide apparaten nog veel aan de aselectheid van de voortgebrachte sleutel. De ontwikkeling van electronische apparatuur zal hierin ongetwijfeld verbetering brengen.

6.3 Gefractioneerde substitutie

Een geheel andere oplossing van het vraagstuk wordt gegeven door de informatie-theorie. Met behulp van een eenvoudig voorbeeld zullen we laten zien welke resultaten bereikt kunnen worden.

Stel dat we een taal bezitten van 4 letters: *A*, *B*, *C* en *D* resp. met de waarschijnlijkheden: *1/2*, *1/4*, *1/8* en *1/8*.

De 4 letters worden vervangen door een binaire substitutie, zodanig dat in een cryptogram de getallen *1* en *0* dezelfde frequentie zullen hebben. Bij de substitutie zal dus rekening gehouden moeten worden met de waarschijnlijkheden waarmee de letters kunnen voorkomen in een tekst.

De kans op de letter *A* is *1/2* en de kans op het voorkomen van de andere letters samen is eveneens *1/2*. We vervangen de *A* door *0* en *B*, *C* en *D* door *1*.

Nu moeten *B*, *C* en *D* nog onderscheiden worden. De kans op *B* is *1/4*, en de kans op *C* en *D* samen is eveneens *1/4*. We voegen daarom aan de substitutie van de *B* een *0* toe, en aan die van de *C* en *D* een *1*. Tenslotte moeten de *C* en *D* nog onderscheiden worden, en dat bereiken we door achter de *C* nog een *0* en achter de *D* een *1* te plaatsen.

A — 0
B — 1 0
C — 1 1 0
D — 1 1 1

We vercijferen nu de volgende zin in de 4-lettertaal:

ADABACABABACADAB

De letters komen voor met de juiste frequentie: 8 *A*'s, 4 *B*'s, 2 *C*'s, 2 *D*'s, in overeenstemming met de waarschijnlijkheden.

A	D	A	B	A	C	A	B	A	B	A	C	A	D	A	B
0	111	0	10	0	110	0	10	0	10	0	110	0	111	0	10

In het cryptogram tellen we 14 nullen en 14 enen. De statistiek van het cryptogram is dus ideaal. Een nadeel is echter dat het cryptogram langer is dan het oorspronkelijke bericht. Het bestaat uit 28 tekens terwijl het bericht 16 letters telt. Deze verlenging is slechts schijn en kan ongedaan worden gemaakt door een tweede substitutie waarin alle binaire combinaties voorkomen:

00 — A
01 — B
10 — C
11 — D

De tweede phase van de gefractioneerde substitutie geeft:

01	11	01	00	11	00	10	01	00	11	00	11	10	10
B	D	B	A	D	A	C	B	A	D	A	D	C	C

Het cryptogram telt nu slechts 14 letters: 4 *A*'s, 3 *B*'s, 3 *C*'s en 4 *D*'s. Het cryptogram is 2 letters korter dan het oorspronkelijke bericht. De statistische kenmerken van de taal zijn in eerste benadering verdwenen.

Ook voor het Nederlands kunnen we substitutieschema samenstellen. Behalve de binaire substituties kunnen we een alphabet van 25 letters vervangen door een vijftallige substitutie, of een 27-letter alphabet door een drietallige substitutie.

Voor ons voorbeeld kiezen we de binaire substitutie. De zeldzame letters *Q*, *X* en *Y* worden uit het alphabet geschrapt en de cijfers *1* t/m *9* worden aan het alphabet toegevoegd.

De afbeelding van de letters op de binaire substituties is niet zo eenvoudig als het voorbeeld in de 4-lettertaal. De waarschijnlijkheden zijn helaas niet zo goed aangepast aan de substitutiemogelijkheden. Verder is aangenomen dat de frequentie van de cijfers tesamen ongeveer *2*% is. De cijfers zijn daarom aan het eind van de tabel geplaatst. Voor het vercijferen van telegrammen die veel cijfers bevatten zou het aanbeveling verdienen de cijfers aan het hoofd van de tabel te plaatsen.

E—000	G—1010	W—111000	2—1111100
N—001	S—10110	B—111001	3—1111101
O—0100	L—10111	J—111010	4—1111110
I—0101	H—11000	Z—111011	5—1111111
A—0110	V—11001	C—1111000	6—11111100
R—0111	K—11010	P—1111001	7—11111101
D—1000	M—110110	F—1111010	8—11111110
T—1001	U—110111	1—1111011	9—11111111

00000—A	01000—I	10000—R	11000—2
00001—B	01001—J	10001—S	11001—3
00010—C	01010—K	10010—T	11010—4
00011—D	01011—L	10011—U	11011—5
00100—E	01100—M	10100—V	11100—6
00101—F	01101—N	10101—W	11101—7
00110—G	01110—O	10110—Z	11110—8
00111—H	01111—P	10111—I	11111—9

We vercijferen als voorbeeld het motto van de Statistische Dag 1954:

MACHT EN ONMACHT VAN DE STATISTIEK

Eerste substitutie:

M	A	C	H	T	E	N	O	N
11011O	011O	1111000	11000	1001	000	001	0100	001

M	A	C	H	T	V	A	N	D
11011O	011O	1111000	11000	1001	11001	011O	001	1000

E	S	T	A	T	I	S	T	I	E
000	10110	1001	011O	1001	0101	10110	1001	0101	000

K
11010

Tweede substitutie:

11011	00110	11110	00110	00100	10000	01010	00011	10110
5	G	8	G	E	R	K	D	Z
01101	11100	01100	01001	11001	01100	01100	00001	01101
N	6	M	J	3	M	M	B	N
00101	10100	10101	10110	10010	10100	01101	0	
F	V	W	Z	T	V	N		

Het cryptogram luidt:

5G8GE RKDZN 6MJ3M MBNFV WZTVN

Om dit cryptografisch systeem met succes te kunnen toepassen zal het gemechaniseerd moeten worden.

6.5 Macht en onmacht in de cryptologie

In de voorgaande paragrafen is aangetoond hoe mathematische methoden gebaseerd op statistische taalkenmerken ons in staat stellen een groot aantal

cryptografische systemen te analyseren. De kennis van de statistische taalstructuur stelt ons bovendien in staat methoden te ontwikkelen waarin de statistische aangrijpingspunten geheel of gedeeltelijk ontbreken.

De statistiek stelt op deze manier een grens aan haar toepassingen. De macht van de statistiek wordt tot eigen onmacht.

Mathematisch Centrum, statistische afdeling.
A. C. Nielsen Co Ltd, Marketing Research, statistische afdeling.

Literatuur

Historisch:

1) E. Bazeries: Les chiffres secrets dévoilés. Etude historique sur les chiffres, appuyée de documents inédits tirés des différents dépôts d'archives (1901).
2) E. Dröscher: Die Methoden der Geheimschriften unter Berucksichtigung ihrer geschichtlichen Entwicklung (1921).
3) A. Meister: Die Geheimschrift im Dienste der päpstlichen Kurie von ihren Anfänge bis zum Ende des XVI Jahrhunderts (1906).
4) F. Pratt: Secret and urgent, the story of codes and ciphers (1939).
5) M. de Vries: Geheimschriften in de Joodse literatuur (1953).

Algemeen:

6) R. Baudouin: Eléments de cryptographie (1939).
7) H. F. Gaines: Elementary cryptanalysis (1939).
8) M. Givierge: Cours de cryptographie. (1936)
9) L. Sacco: Manuelle de cryptographie (1951).

Informatie-theorie:

10) C. E. Shannon: Communication theory of secrecy systems (1949).
11) M. de Vries: Concealment of information (1953) (nog niet gepubliceerd).
12) A. van Wijngaarden: Informatie en communicatie (nog niet gepubliceerd).