Rを用いたテキストマイニング６
—頭韻のマーキング—

岩根　久

大阪大学サイバーメディアセンター
〒560-0043 豊中市待兼山町 1–32
Email: iwane@les.cmc.osaka-u.ac.jp

概要　フランス詩において、脚韻という技法が広く用いられているが、時に近接する単語の中で同じ子音を繰り返して、音の効果を狙う頭韻という技法が用いられることがある。本稿では、１６世紀フランスの詩人ロンサール（1524-1585）の『恋愛詩集』（*Les Amours*, 1552）を例に、頭韻の技法が顕著に用いられている箇所を明らかにし、視覚化する手法を概説する。

キーワード　R, テキストマイニング, フランス詩、頭韻

Text Mining using R (6): Method for Marking Alliteration

Hisshi IWANE

Cybermedia Center, Osaka University
1-32 Machikaneyama-cho, Toyonaka, Osaka, 560-0043 Japan

Abstract　French poets commonly use a rhetorical device called rhyme. Sometimes they use another device called alliteration, repeating the same consonant in the words in close proximity to each other to obtain a sound effect. Taking as an example *Les Amours*(1552) of Ronsard(1524-84), this paper briefly explains how to find passages in which the device of alliteration is prominently used

Keywords　alliteration, French poetry, OpenOffice, R, text mining,

1. はじめに

本稿執筆者は、文学研究を補完するための手段として、これまでに様々な計量的な手法を提案してきた[1]。本稿では、その一環として、１６世紀フランスの詩人ロンサール（1524-1585）の『恋

[1]岩根久,「フランス詩のマイニング—道具のR、Rの道具—」,『統計数理研究所共同研究リポート 231：多変量アプローチによるテキストの計量研究』, 統計数理研究所, 2009 年 3 月, pp. 23-52.、「Rを用いたテキストマイニング１—文字出現頻度によるソネのマイニング—」,『言語文化共同研究プロジェクト 2008：電子化言語資料分析研究　2008-2009』, 大阪大学大学院言語文化研究科, 2009 年 5 月, pp. 25-34、「Rを用いたテキストマイニング２—ソネの文字構成について—」,『言語文化共同研究プロジェクト 2009：電子化言語資料分析研究　2009-2010』, 大阪大学大学院言語文化研究科, 2010 年 5 月, pp. 29-40、「Rを用いたテキストマイニング３—叙事

愛詩集』（*Les Amours*, 1552）を例に、R[2]を用いて、頭韻の技法が顕著に用いられている箇所を明らかにする方法、また、そのデータに基づき Apache OpenOffice[3]を用いて視覚的に表示する方法を提案し、必要に応じて R スクリプトの実例を示しながら概説する。

2. 着想および目的

ロンサールの『恋愛詩集』（*Les Amours*, 1552）（以下、A52 と略称）は、序詩の 14 行詩 1 編および 182 篇の 14 行詩からなるソネ集である。ロンサールはこのソネ集により、フランスのソネの標準的な押韻構成を確立したとされている。以下の詩句は、のソネ第 1 番からの引用である。

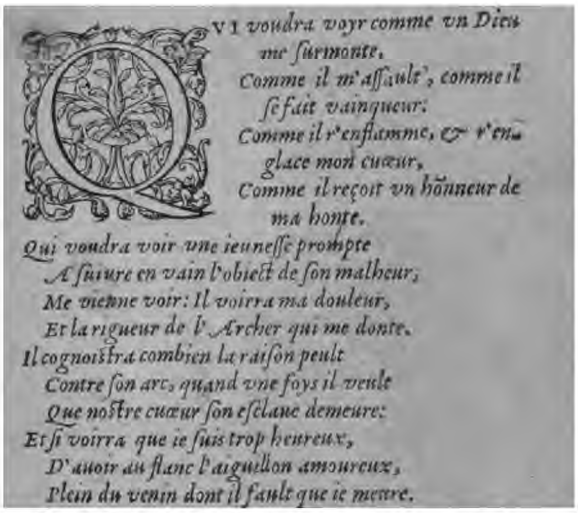
VI voudra voyr comme vn Dieu
me ſurmonte,
Comme il m'aſſault, comme il
ſe fait vainqueur:
Comme il s'enflamme, & s'en-
glace mon cueur,
Comme il reçoit vn hõneur de
ma honte.
Qui voudra voir vne ieuneſſe prompte
A ſuiure en vain l'obiect de ſon malheur,
Me vienne voir: Il voirra ma douleur,
Et la rigueur de l'Archer qui me donte.
Il cognoiſtra combien la raiſon peult
Contre ſon arc, quand vne foys il veult
Que noſtre cueur ſon eſclaue demeure:
Et ſi voirra que ie ſuis trop heureux,
D'auoir au flanc l'aiguillon amoureux,
Plein du venin dont il fault que ie meure.

図 1 Ronsard, *Les Amours*, Sonnet 1

詩と抒情詩の計量的差異―」，『言語文化共同研究プロジェクト 2010：電子化言語資料分析研究　2010-2011』，大阪大学大学院言語文化研究科, 2011 年 5 月, pp. 23-30.、「押韻ラベル付与の自動化に向けて―ソネの押韻構成分析のためのRスクリプト―」,『統計数理研究所共同研究リポート 278：マイニング技術を応用したテクスト分析研究』, 統計数理研究所, 2012 年 3 月, pp. 17-22、「Rを用いたテキストマイニング４―ソネ脚韻の音素分析―」，『言語文化共同研究プロジェクト 2011：電子化言語資料分析研究　2011-2012』, 大阪大学大学院言語文化研究科, 2012 年 5 月, pp. 29-39、「脚韻判定法の評価―フランスのソネの押韻構成を用いて―」，『統計数理研究所共同研究リポート 298：統計的マイニング技術を応用したテクスト研究』, 統計数理研究所, pp. 49-60, 2013 年 3 月、「イタリアソネットの自動押韻ラベル付与―ペトラルカ『カンツォニエーレ』のソネット―」，『統計数理研究所共同研究リポート 322：人文学データのマイニング』, 統計数理研究所, 2014 年 3 月, pp. 25-33、「１６世紀フランス詩の音の分析に向けて―N-gram とデンドログラムの利用―」,『統計数理研究所共同研究リポート 345：人文学データのマイニング II』・統計数理研究所, 2015 年 3 月, pp.11-20.、「Rを用いたテキストマイニング５―N-gram による１６世紀フランス詩の分析―」,『言語文化共同研究プロジェクト 2014：電子化言語資料分析研究　2014-2015』, 大阪大学大学院言語文化研究科, 2015 年 5 月, pp.15-22.、「コーパス分析ソフト CasualConc の活用(1)―ロンサールの論説詩とプロテスタント詩人による版ロンサール詩を例として―」,『ロンサール研究』, 30 号, 日本ロンサール学会, 2017 年 8 月, pp.139-145.、「コーパス分析ソフト CasualConc の活用(2)―例として『恋愛詩集』(1552)を中心に―」，『ロンサール研究』, 31 号, 日本ロンサール学会, 2018 年 8 月, pp.99-104.

[2] http://www.r-project.org/ 参照。OSはWindows 10 pro、Rは R version 4.0.2 (x86_64-w64-mingw32/x64 (64-bit))を用いた。

[3] http://www.openoffice.org/ 参照。Apache OpenOffice 4.1.7を用いた。

Qui voudra voyr comme un Dieu me surmonte,
Comme il m'assault, comme il se fait vainqueur,
Comme il r'enflamme, & r'englace mon cuoeur,
Comme il reçoit un honneur de ma honte,
Qui voudra voir une jeunesse prompte
A suyvre en vain l'object de son malheur,
Me vienne voir: il voirra ma douleur,
Et la rigueur de l'Archer qui me donte.
Il cognoistra combien la raison peult
Contre son arc, quand une foys il veult
Que nostre cuoeur son esclave demeure:
Et si voirra que je suis trop heureux,
D'avoir au flanc l'aiguillon amoureux,
Plein du venin dont il fault que je meure.

ひとりの神がいかに私を打ちのめし
いかに私を攻め　いかに勝ち誇るか
いかに私を恋の炎で燃え上がらせては　凍りつかせるか
いかに私の恥辱をもって身の誉とするのか見たい者は、
ひとりの若者が　自らを不幸にするひとを
空しいとは知りつつも　つい追いかけてしまうのを見たい者は、
私を見に来るがよい。そうすれば私の苦しみと
私を意のままにするあの射手の酷さを目にするだろう。
そして知るだろう。分別などその弓にどれほど
抗えるものか、ひとたび射手が
われらの心を虜にしたいと望むなら。
そして目にするだろう。脇腹には死に至る
恋（アモール）の毒矢を受けているのに
こんなに幸せな私を。[4]

引用のフランス語テキストで灰色の網掛けにしたcとvの文字は、一見して繰り返されていることが感じられる子音である。音声的には、qも含めて[k]の音と[v]の音が詩句の中に縦横に散りばめられていることが見て取れる。音声的効果という意味では、詩句を音声記号に変換し、それを分析するのが正当な方法であるが、16世紀フランス語の文字テキストを自動的に音声記号に変換する手法はまだ確立していないため、本稿では暫定的に文字情報のみを利用して分析する。

ここで、文字使用の頻度の指標としてz値（平均値からの距離を標準偏差の倍数で表した数値）を用いる。たとえば、cに関しては、ソネ第1番の語頭のcの文字頻度は10、全ソネ183篇で各ソネあたりの語頭のcの文字頻度の平均値は6.1、標準偏差は2.56なので、z値は1.5となる。vに関しては、同様に計算して、z値は3.8となる。

このように、各ソネにおける語頭文字のz値を求めることにより、それぞれのソネで頻用されている文字を割り出すことができる。表1は、序詩のソネ（S000）からソネ第9番（S009）までの各ソネそれぞれで頻用される語頭文字を示したものである。頻用の程度を、1＜z≦2、2＜z≦3、3＜zに分類している。3＜zが最も程度の高い頻用を示す。ただし、S004のyのように、も

[4] 図1は、フランス国立図書館所蔵のデジタル化資料より：Ronsard, Pierre de, Les Amours ..., Paris, Vve M. de La Porte, 1552. Bibliothèque nationale de France, département Réserve des livres rares, RES P-YE-1482/NUM M- 1040604. 日本語訳は現在進行中の日本ロンサール学会有志グループの訳による。

ともと出現頻度の低い語頭文字（平均値 0.5、標準偏差 0.68）は、頻度が低くても（S004 では頻度 2）z 値が高くなるので注意を要する。

表 1　同一ソネで頻用される語頭文字

z 値	S000	S001	S002	S003	S004	S005	S006	S007	S008	S009
1＜ z ≦2	is	chru	a		cnt	cel	g	gimnpt	dt	dsu
2＜ z ≦3	a		de	f	y	h				r
3＜ z		iv					c			

以下、すべてのソネについて表 1 のような資料を作成するための R スクリプトを示す。

3.　同一ソネで頻用される語頭文字データ抽出のための R スクリプトの概要

・処理の対象となる A52 のテキストファイル a52.txt を作業ディレクトリに置き、ロケールをフランスにした上で、エンコーディング UTF-8 で、R のデータフレーム変数に読み込む。

```
Sys.setlocale(,"fra")
a52<-read.delim("a52.txt",row.names=1,encoding="UTF-8")
```

a52.txt の冒頭の 5 行は以下の通り。

	WN	SN	VN	TXT
1	A52	S000	01	Divin troupeau,qui sur les rives molles
2	A52	S000	02	Du fleuve Eurote, ou sur le mont natal,
3	A52	S000	03	Ou sur le bord du chevalin crystal,
4	A52	S000	04	Assis, tenez vos plus sainctes escolles:
5	A52	S000	05	Si quelque foys aux saultz de vos carolles

左端の列は通し番号、各列は、WN（作品ラベル）、SN（ソネ番号）、VN（詩句番号）、TXT（詩句テキスト）に対応している。各列は TAB コードで区切られている。

・綴り字記号を取り除く関数の定義を読み込む。元のテキストから綴り字記号を取り除き、処理を単純化するためである。

```
StripAcc<-function(x){
#Character with Accent to Normal Character
chartr("àâéèêëîïôòûùüç","aaeeeeiiiioooouuuuc",x)}
```

・引数のテキストに出現する単語の語頭文字をカウントする関数の定義を読み込む。

```
WordTopCharCount<-
function(x){table(factor(substring(unlist(strsplit(tolower(StripAcc(x)),"[ ,.?!;:]+")),1,1),
levels=letters))}
```

この関数はテキストから綴り字記号を取り除き、小文字に統一した上で、単語に分解し、その単語の第 1 文字を取り出し、アルファベット各文字の集計表を作成する。

・ソネごとの語頭文字をカウントした結果を読み込む。

```
WTCS<-sapply(split(a52$TXT,a52$SN),WordTopCharCount)
```

結果を読み込んだ変数 WTCS は頻度行列（行にソネ番号、列にアルファベット）である。

各語頭文字ごとの頻度分布を図 2 に示す。k, w で始まるような単語は A52 には存在しないので、その頻度は 0 となっている。

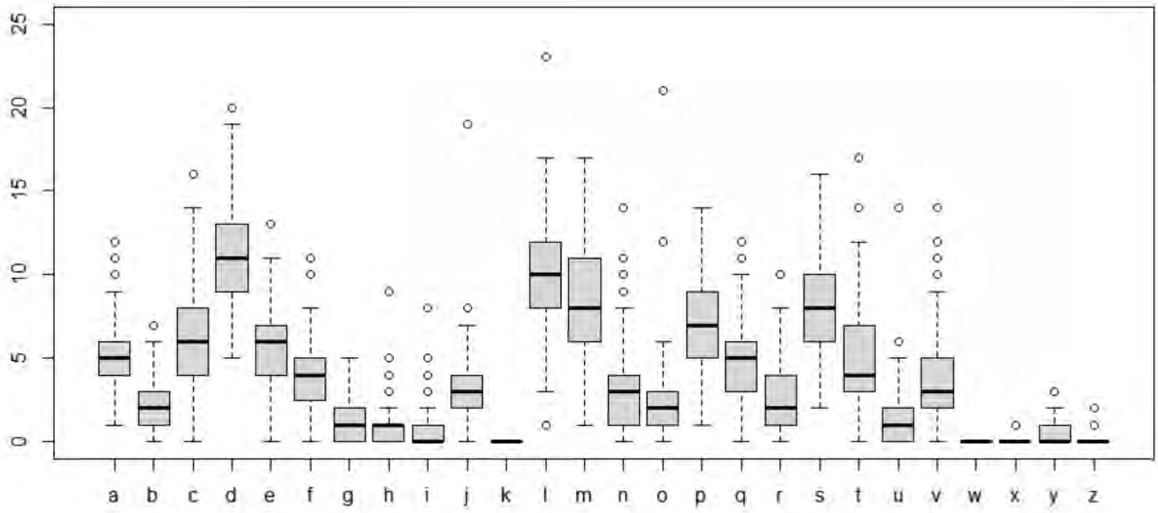

図 2 語頭各文字の全ソネにおける頻度分布を示す箱ひげ図

・語頭文字の頻度の各ソネあたりの平均値と標準偏差を求める。

```
ssd<-function(x){sqrt(var(x)*(length(x)-1)/length(x))}
WTCS_M<-apply(WTCS,1,mean)
WTCS_SD<-apply(WTCS,1,ssd)
```

R に組込みの関数 sd は不偏分散に基づく標準偏差を求める関数なので、標本分散に基づく標準偏差を求める関数 ssd を定義する。アルファベット各文字の語頭頻度の平均値を WTCS_M、標準偏差を WTCS_SD に格納する。

・各ソネにおける高頻度の語頭文字を z 値に従って抽出する。

```
FU<-(WTCS>(WTCS_M+WTCS_SD))
FFU<-(WTCS>(WTCS_M+2*WTCS_SD))
FFFU<-(WTCS>(WTCS_M+3*WTCS_SD))

FU1<-xor(FU,FFU)
FU2<-xor(FFU,FFFU)
FU3<-FFFU

CFU1<-apply(FU1,2,function(x){paste(letters[x],collapse="")})
CFU2<-apply(FU2,2,function(x){paste(letters[x],collapse="")})
CFU3<-apply(FU3,2,function(x){paste(letters[x],collapse="")})
```

FU は頻度行列 WTCS の各要素が z ＞1 となるかどうかの真偽値の行列、FFU は z ＞2、FFFU は z ＞3 となるかどうかの真偽値の行列である。FU1、FU2、FU3 はそれぞれ 1＜ z ≦2、2＜ z ≦3、3＜ z の範囲にあるかどうかの真偽値の行列である。FU1、FU2、FU3 を用いて、アルファベットから文字を抽出したものが、CFU1、CFU2、CFU3 となる。

CFU1、CFU2、CFU3 を結合すれば、表 1 のような資料ができる。図 3 は、CFU1、CFU2、CFU3 を結合して、画面に出力した結果の一部である。

```
> rbind(CFU1,CFU2,CFU3)
     S000 S001   S002 S003 S004  S005  S006 S007     S008 S009  S010   S011 S012
CFU1 "is" "chru" "a"  ""   "cnt" "cel" "g"  "gimnpt" "dt" "dsu" "dfjn" "p"  "cr"
CFU2 "a"  ""     "de" "f"  "y"   "h"   ""   ""       ""   "r"   ""     "m"  "p"
CFU3 ""   "iv"   ""   ""   ""    ""    "c"  ""       ""   ""    ""     ""   "j"
     S013  S014    S015   S016 S017   S018  S019 S020   S021   S022  S023  S024
CFU1 "dru" "jmnpt" "chlu" "pv" "bfmu" "dfo" "fs" "gjuv" "ipqv" "fsu" "ors" ""
CFU2 ""    ""      ""     "em" ""     ""    ""   "b"    ""     "ay"  "e"   "y"
CFU3 ""    ""      "z"    "z"  ""     "u"   "t"  ""     ""     ""    "cz"  "r"
     S025 S026    S027 S028   S029   S030  S031 S032 S033  S034  S035  S036  S037
CFU1 "em" "jlotv" "bf" "efip" "fjms" "dmu" "d"  ""   "ftv" "djp" "cqs" "dlt" "co"
CFU2 "ny" "by"    ""   "n"    ""     ""    ""   "s"  ""    "m"   "i"   "p"   ""
CFU3 ""   ""      ""   ""     ""     ""    ""   ""   ""    ""    ""    "x"   ""
     S038 S039  S040   S041 S042 S043 S044  S045 S046 S047 S048   S049   S050
CFU1 "s"  "rs"  "fops" "ho" "ls" "p"  "apv" "s"  ""   "hm" "acls" "dosu" "l"
CFU2 "d"  "ben" ""     ""   "ai" "t"  ""    "d"  "d"  "q"  "r"    ""     "dg"
CFU3 ""   ""    ""     ""   ""   ""   ""    ""   "o"  ""   ""     ""     "nz"
```

図 3 同一ソネで頻用される語頭文字の出力画面

すべてのソネについて、同一ソネで頻用される語頭文字に関するデータは、末尾の表 2 に掲載する。

4. Apache OpenOffice による同一ソネで頻用される語頭文字の可視化

Amour, amour, donne moy paix ou trefve,
Ou bien retire, & d'un garrot plus fort
Tranche ma vie, & m'avance la mort,
Me bienheurant d'une langueur plus bréve.
Soit que le jour ou se couche, ou se leve,
Je sens tousjours un penser qui me mord,
Et contumax au cours de son effort,
De pis en pis mes angoisses r'engreve.
Que doibs je faire? Amour me faict errer,
Si haultement que je n'ose esperer
De mon salut que la desesperance.
Puis qu'Amour donc ne me veult secourir,
Pour me deffendre il me plaist de mourir,
Et par la mort trouver ma delivrance.

たとえば、上に引用したの第１１番目のソネは、頻度の高い語頭文字 m（2＜ z ≦3）と p（2＜ z ≦3）を含んでいることがわかっていれば、下記のように自動でタグ付けすることが可能である（ここの本題ではないので、スクリプトは省略する）。

Amour, amour, donne <2m>oy <1p>aix ou trefve,
Ou bien retire, & d'un garrot <1p>lus fort

Tranche <2m>a vie, & <2m>'avance la <2m>ort,
<2M>e bienheurant d'une langueur <1p>lus bréve.
Soit que le jour ou se couche, ou se leve,
Je sens tousjours un <1p>enser qui <2m>e <2m>ord,
Et contumax au cours de son effort,
De <1p>is en <1p>is <2m>es angoisses r'engreve.
Que doibs je faire? Amour <2m>e faict errer,
Si haultement que je n'ose esperer
De <2m>on salut que la desesperance.
<1P>uis qu'Amour donc ne <2m>e veult secourir,
<1P>our <2m>e deffendre il <2m>e <1p>laist de <2m>ourir,
Et <1p>ar la <2m>ort trouver <2m>a delivrance.

このようにタグ付けされた、<2...>の部分や、<1...>の部分に違う色付けの書式を設定したい場合は、どのようにすればいいのだろうか。

もちろん、Microsoft Word のワイルドカードによる置換を用いて、これを実行することは可能であるが、ここでは Apache OpenOffice の Writer を用いる方法を紹介したい。

図 4 OpenOffice Writerの正規表現置換

図 4 は、<2...>とタグ付された箇所を検索し、語頭文字部分のみ灰色の網掛け文字に置換した状態を示している。このように、OpenOffice の Writer の正規表現置換を用いれば、タグ付けした箇所を効率的に置換することができる。また、検索文字列、置換文字列に属性や書式を設定することができる。

時に、R などのスクリプト言語を用いてテキストにタグ付けしたあと、タグ付けした箇所に何らかの書式を設定したい、あるいは逆に書式設定されている箇所をタグ付けされたテキストに変換したいなどのケースが生ずる。OpenOffice はこうしたツールのひとつとして有益であろう。

表 2　同一ソネで頻用される語頭文字

z 値	S000	S001	S002	S003	S004	S005	S006	S007	S008	S009
1＜ z ≦2	is	chru	a		cnt	cel	g	gimnpt	dt	dsu
2＜ z ≦3	a		de	f	y	h				r
3＜ z		iv					c			
z 値	S010	S011	S012	S013	S014	S015	S016	S017	S018	S019
1＜ z ≦2	dfjn	p	cr	dru	jmnpt	chlu	pv	bfmu	dfo	fs
2＜ z ≦3		m	p				em			
3＜ z			j			z	z		u	t
z 値	S020	S021	S022	S023	S024	S025	S026	S027	S028	S029
1＜ z ≦2	gjuv	ipqv	fsu	ors		em	jlotv	bf	efip	fjms
2＜ z ≦3	b		ay	e	y	ny	by		n	
3＜ z				cz	r					
z 値	S030	S031	S032	S033	S034	S035	S036	S037	S038	S039
1＜ z ≦2	dmu	d		ftv	djp	cqs	dlt	co	s	rs
2＜ z ≦3			s		m	i	p		d	ben
3＜ z							x			
z 値	S040	S041	S042	S043	S044	S045	S046	S047	S048	S049
1＜ z ≦2	fops	ho	ls	p	apv	s		hm	acls	dosu
2＜ z ≦3			ai	t		d	d	q	r	
3＜ z							o			
z 値	S050	S051	S052	S053	S054	S055	S056	S057	S058	S059
1＜ z ≦2	l	cduv		g	cnr	gjt	eq	fr	iq	aosv
2＜ z ≦3	dg	y	ay	iv		b	b	v		y
3＜ z	nz			h						
z 値	S060	S061	S062	S063	S064	S065	S066	S067	S068	S069
1＜ z ≦2	cdi	jpq	aip	tv		fp	j	bcfo	d	cg
2＜ z ≦3				m	hv					s
3＜ z		v		y				n		y
z 値	S070	S071	S072	S073	S074	S075	S076	S077	S078	S079
1＜ z ≦2	anr	eotv	it	ei	cu	mt	ceo	gh	s	it
2＜ z ≦3		iy	n		h	u	qs			g
3＜ z									e	

z 値	S080	S081	S082	S083	S084	S085	S086	S087	S088	S089
1＜ z ≦2	ist	mru	fs	cfim	p	i	c	bgp	ah	mu
2＜ z ≦3			jm	v			m	y		l
3＜ z	f								g	
z 値	S090	S091	S092	S093	S094	S095	S096	S097	S098	S099
1＜ z ≦2	cp	cops	i	g	cn	bcn		mpt	co	au
2＜ z ≦3			t	b	b		p	y	p	
3＜ z		z	z		f					
z 値	S100	S101	S102	S103	S104	S105	S106	S107	S108	S109
1＜ z ≦2	g	fmr		eit	ns	st	dlpt		befglq	gt
2＜ z ≦3	d		dp					s		
3＜ z			a	fz	q				h	
z 値	S110	S111	S112	S113	S114	S115	S116	S117	S118	S119
1＜ z ≦2	c	m	cfv	fimu	d	asu	et	n	dpt	cel
2＜ z ≦3	d		ny					e	fs	qt
3＜ z			z							
z 値	S120	S121	S122	S123	S124	S125	S126	S127	S128	S129
1＜ z ≦2	fglosu	mp	n	ds	ms	mu	ejm	or	am	aer
2＜ z ≦3			r	gt	j	f	p			
3＜ z						i				
z 値	S130	S131	S132	S133	S134	S135	S136	S137	S138	S139
1＜ z ≦2		lps	h	qt	a	pu	a	p		c
2＜ z ≦3	cej	i			b				t	b
3＜ z		v								
z 値	S140	S141	S142	S143	S144	S145	S146	S147	S148	S149
1＜ z ≦2	chs	flq	nt	befm	ef	fv	dm	elnt	a	rs
2＜ z ≦3					l		j			
3＜ z		g		n		o			glu	
z 値	S150	S151	S152	S153	S154	S155	S156	S157	S158	S159
1＜ z ≦2	cl	d	bm	mq	fis	afhstv	bes	pq	bm	iq
2＜ z ≦3			c	y					i	bd
3＜ z										

z 値	S160	S161	S162	S163	S164	S165	S166	S167	S168	S169
1＜ z ≦2	abehru	a	gn	i	t	diu	d	jp	alsu	as
2＜ z ≦3				m	my		v			
3＜ z					r					
z 値	S170	S171	S172	S173	S174	S175	S176	S177	S178	S179
1＜ z ≦2	fu	qs	r	at	bm	lt	fptv	ijqs	ept	env
2＜ z ≦3	s			h		a		b		
3＜ z										

z 値	S180	S181	S182
1＜ z ≦2	jlv	jv	bv
2＜ z ≦3		hq	d
3＜ z			