テキスト マイニングの諸相と フリーソフト の活用

石田 基広

徳島大学総合科学部

テキスト マイニングの諸相

- 日本語テキスト
  - 上田太一郎『 事例で学ぶテキスト マイニング』
  - 大塚&乾&奥村『 意見分析エンジン』
  - 藤井&小杉&李『 福祉・ 心理・ 看護のテキスト マイニング 入門』
  - 林 俊克『 Excel で学ぶテキスト マイニング 入門』
  - 那須川哲哉『 テキスト マイニングを使う 技術/作る 技術』

テキスト マイニングの諸相

- 日本語テキスト
  - 上田太一郎『 事例で学ぶテキスト マイニング』
  - 大塚&乾&奥村『 意見分析エンジン』
  - 藤井&小杉&李『 福祉・ 心理・ 看護のテキスト マイニング 入門』
  - 林 俊克『 Excel で学ぶテキスト マイニング 入門』
  - 那須川哲哉『 テキスト マイニングを使う 技術/作る 技術』
- 応用事例
  - 公開特許文書を利用した技術傾向の把握
  - 社説タイト ルを利用した社会動向の把握
  - キャラクタグッズの商品化調査
  - 株のネットト レード に関するアンケート 調査

技術としてのテキスト マイニング

- 大量のテキスト 集合をデータ 分析のインプットとするための技術
  - テキスト 処理のフェーズ (自然言語処理)
  - データ 処理のフェーズ (データ マイニング)

技術としてのテキスト マイニング

- 大量のテキスト 集合をデータ分析のインプットとするための技術
  - テキスト 処理のフェーズ (自然言語処理)
  - データ処理のフェーズ (データマイニング)
- 「 テキスト 」の種類
  - 公開された文書群
    - テキスト の類似性，分類などの課題
    - Baldi 他『 確率モデルによる Web データ 解析法』

技術としてのテキスト マイニング

- 大量のテキスト 集合をデータ分析のインプットとするための技術
  - テキスト 処理のフェーズ (自然言語処理)
  - データ処理のフェーズ (データマイニング)
- 「 テキスト 」の種類
  - 公開された文書群
    - テキスト の類似性，分類などの課題
    - Baldi 他『 確率モデルによる Web データ 解析法』
  - 調査アンケート (自由記述文)
    - 蓄積された顧客意見の分析 (企業など)
    - 自由記述形式のアンケート 分析 (大学研究者など)

テキスト マイニングのツール

市販のツール

- Text Mining for Clementine
- True Teller
- Text Mining Studio
- Text Miner
- Trustia

テキスト マイニングのツール

市販のツール

- Text Mining for Clementine
- True Teller
- Text Mining Studio
- Text Miner
- Trustia

本日のお題: アンケート 調査への応用事例

テキスト マイニングのツール

市販のツール

- Text Mining for Clementine
- True Teller
- Text Mining Studio
- Text Miner
- Trustia

本日のお題: アンケート 調査への応用事例

フリーの解析ツール

- 自然言語処理 和布蕪，南瓜
- データマイニング **R**

テキスト マイニングのツール

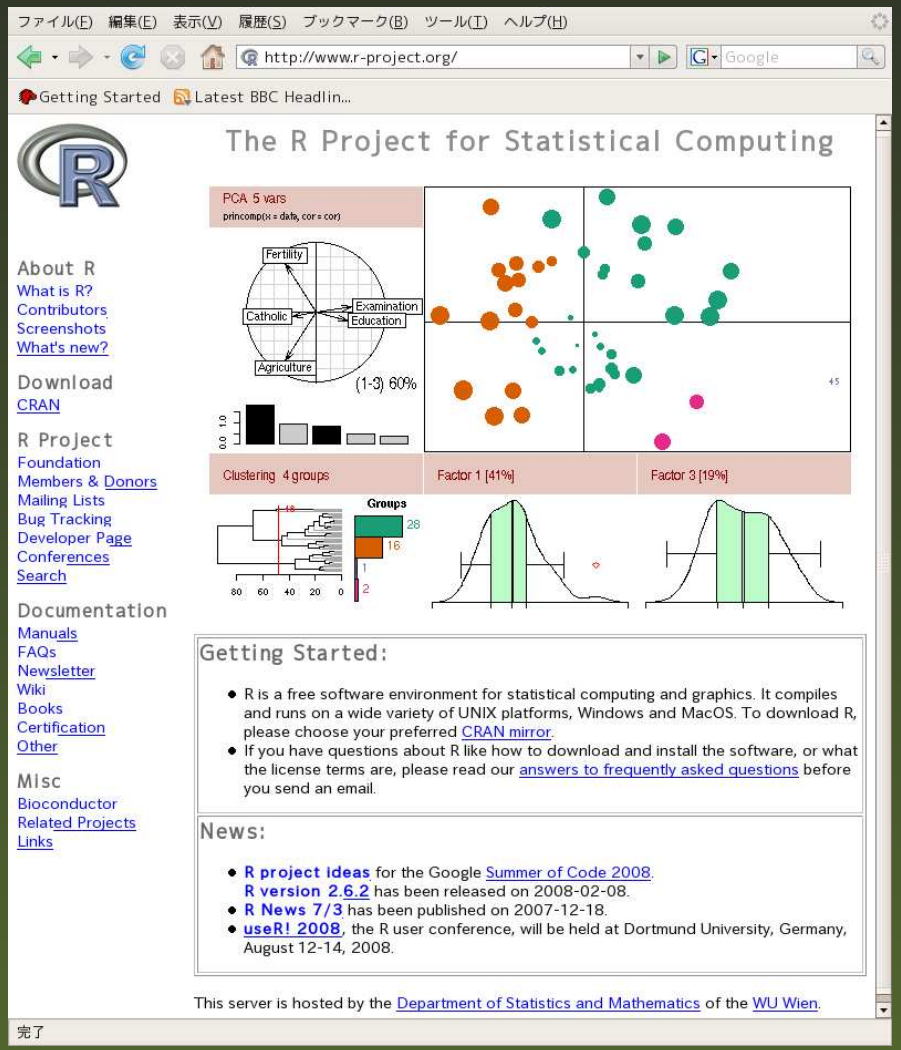
市販のツール

- Text Mining for Clementine
- True Teller
- Text Mining Studio
- Text Miner
- Trustia

本日のお題: アンケート 調査への応用事例

フリーの解析ツール

- 自然言語処理 和布蕪，南瓜
- データマイニング **R**



アンケート 分析への応用事例

- 2007 年度科学研究費:
  「 日本語の対人配慮表現の多様性」
  (代表: 大阪府立大学 野田尚史)
  - 「 ことばに関するアンケート 」
    (徳島大学: 岸江信介)
  - 日本語文章を書いてもらう 設問 16 個
  - 総回答者数 286 名

アンケート 分析への応用事例

- 2007 年度科学研究費:
  「 日本語の対人配慮表現の多様性」
  (代表: 大阪府立大学 野田尚史)
  - 「 ことばに関するアンケート 」
    (徳島大学: 岸江信介)
  - 日本語文章を書いてもらう 設問 16 個
  - 総回答者数 286 名

「 このアンケート 調査は、みなさんがふだん使っている話しことばの使い方について全国的な地域差や世代差について調べることを目的としています。」

「ことばに関するアンケート」

- 問 3 自宅の近くにある名所に友達と久しぶりに行きました。一緒に写真をとろうということになったのですが、シャッターを誰かに押してもらわなければなりません。そうすると、ちょうどタイミングよく、近くに3 人の中学生のグループがいました。その時、その中の一人の男の子に頼んで写真をとってもらおうと思ったとします。
- (1) さて、実際に写真をとってもらうように頼みますか。いずれかを選んで番号を右の回答欄に記入してください。

1 頼む　→ (2) へ　　・　　2 頼まない　→ (3) へ

- (2)「頼む」と答えた方のみにお聞きします。この時、この中学生の一人の男の子に「頼む」とすればどう言って頼むか、そのセリフを書いてください。

- (3) この場面では、この中学生の一人の男の子に頼むのにどの程度気をつかうか、その度合いについて下から適当なものを選び、右の回答欄に番号を記入して下さい。

1. 非常に気をつかう 2. かなり気をつかう 3. 少しは気をつかう 4. あまり気をつかわない 5. まったく気をつかわない

- 問 4 友達と一緒に写真をとってほしいと思っていたら、ちょうど近くに中年の夫婦がいました。その時、男性の方に頼んで写真をとってもらおうと思ったとします。

被験者の情報 問 17

- 「 あなたが小中学校の大半をお過ごしになった地域を教えてください」
- 「 現在、住んでいる住所を教えてください」
- 「 世代について教えて下さい」
- 「 性別についてお伺いします」
- 「 ご職業」

性別	世代	東日本 (E)	西日本 (W)
M	10 代	2	17
	20 代	7	35
F	10 代	20	49
	20 代	38	114
	30 代	1	0
	40 代	0	1
	50 代	0	2

問 3(2) – 以下 Q3A2 と略 –, 問 4(2) –以下 Q4A2 と略 – の両方に自由記述文を残したのは 157 名．なお 問 3(3), 問 4(3) は遠慮の程度を尋ねて，それぞれ Q3A3, Q4A3 と略

R によるデータ分析

- 遠慮の度合いについてのクロス表
-

xtabs(~Q3A3 + Q4A3)

Q4A3 (中年男性)	遠慮の程度	H(高)	<–>	L(低)	
Q3A3 (中学生)	H1	H2	M3	L4	L5
H1	10	1	2	1	0
H2	15	30	14	1	0
M3	7	47	58	3	0
L4	2	11	17	10	1
L5	1	1	2	0	1

R によるデータ分析

- 遠慮の度合いについてのクロス表

xtabs(~Q3A3 + Q4A3)

Q4A3 (中年男性)	遠慮の程度	H(高)	<–>	L(低)	
Q3A3 (中学生)	H1	H2	M3	L4	L5
H1	10	1	2	1	0
H2	15	30	14	1	0
M3	7	47	58	3	0
L4	2	11	17	10	1
L5	1	1	2	0	1

- R によるマクネマーの検定

```
mcnemar.test (xtabs(~Q3A3 + Q4A3))
McNemar's chi-squared = 56.3469, df = 10, p-value = 1.768e-08
```

自由記述欄への回答例

ID	EW	Age	Sex	Q3A3
7	W	20 代	F	M3
すみません。写真撮ってくれませんか？				
17	W	20 代	F	H2
すいません、ちょっと写真を撮ってもらいたいんですけど、いいですか？				
27	E	20 代	F	M3
すみません。写真を撮って頂けますか？				
32	W	10 代	M	M3
ちょっとごめん。写真撮ってくれん？				
37	W	10 代	F	M3
写真を撮ってもらいたいんだけど、撮ってもらえないかな？				

欧米語テキスト の解析

- **tm** パッケージ (tm_0.2-3 by Ingo Feinerer)
  - 各種フォーマット・メタ情報 (HTML,XML,Gmane,RSS) への対応
  - stopwords (and, or, ...) の削除処理 (英独露など 13 言語に対応)
  - stemming (study -> studies, studied, studying) の処理 (11 言語に対応)
  - 文章・ターム行列の作成，各種重み付け (tf-idf など)

欧米語テキスト の解析

- **tm** パッケージ (tm_0.2-3 by Ingo Feinerer)
  - 各種フォーマット・メタ情報 (HTML,XML,Gmane,RSS) への対応
  - stopwords (and, or, ...) の削除処理 (英独露など 13 言語に対応)
  - stemming (study -> studies, studied, studying) の処理 (11 言語に対応)
  - 文章・ターム行列の作成，各種重み付け (tf-idf など)

```
library(tm) # Text Mining パッケージ
file.obj <- "/Target/textDir"
# 内容は Alice was beginning to get very tired of sitting by her sister on the bank,..
alice.DC <- TextDocCol(DirSource(file.obj),
        readerControl = list(reader = readPlain, language = “en_US”, load = TRUE ))
alice.DC # show(alice.DC)
A text document collection with 1 text document
# stopWords を削除
alice.DC3 <- tmMap(alice.DC2, removeWords, stopwords("english"))
# Stemming を行う
alice.DC4 <- tmMap(alice.DC3, stemDoc)
inspect(alice.DC4) # 結果を表示
 "alic" "was" "begin" "to" "get" "veri" "tire" "of" "sit" "by" "her" "sister" "on" "the" “bank” ...
```

和布蕪による 日本語形態素解析

```
$ mecab          # 形態素解析を行う
すみません。写真撮ってくれませんか？
```

和布蕪による日本語形態素解析

```
$ mecab          # 形態素解析を行う
すみません。写真撮ってくれませんか？
```

```
すみません    感動詞,*,*,*,*,*, すみません, スミ マセン, スミ マセン
。            記号, 句点,*,*,*,*,。 ,。 ,。
写真          名詞, 一般,*,*,*,*, 写真, シャシン, シャシン
撮っ          動詞, 自立,*,*, 五段・ラ行, 連用タ接続, 撮る, トッ, トッ
て            助詞, 接続助詞,*,*,*,*, て, テ, テ
くれ          動詞, 非自立,*,*, 一段・クレル, 連用形, くれる, クレ, クレ
ませ          助動詞,*,*,*, 特殊・マス, 未然形, ます, マセ, マセ
ん            助動詞,*,*,*, 不変化型, 基本形, ん, ン, ン
か            助詞, 副助詞／並立助詞／終助詞,*,*,*,*, か, カ, カ
？            記号, 一般,*,*,*,*, ？, ？, ？
EOS
```

和布蕪を R から 使う

和布蕪ライブラリ について： http://mecab.sourceforge.net/libmecab.html

R の C 言語ライブラリ と 和布蕪 の C ライブラリ の結合

和布蕪を R から 使う

和布蕪ライブラリ について： http://mecab.sourceforge.net/libmecab.html

R の C 言語ライブラリ と 和布蕪 の C ライブラリ の結合

```
# ごく 単純な実装例 RMeCab パッ ケージ？
#include <R.h>
#include <Rdefines.h>
#include <mecab.h>
#include <stdio.h>
SEXP mecab3(SEXP aa){
        SEXP parsed;
        const char* input = CHAR(STRING_ELT(aa,0));
        mecab = mecab_new2 (input);
        CHECK(mecab);
       result = mecab_sparse_tostr(mecab, input);
        CHECK(result);
        PROTECT(parsed = mkString(result));
        UNPROTECT(1);
        mecab_destroy(mecab);
        return(parsed);
}
```

RMeCab パッケージによる解析

```
res <- .Call ("myMecab",” 写真撮ってくれませんか。”)
 名詞      動詞     助詞     動詞     助動詞    助動詞    助詞
"写真"    "撮っ"    "て"    "くれ"    "ませ"      "ん"      "か"
```

RMeCab パッケージによる解析

```
res <- .Call ("myMecab",” 写真撮ってくれませんか。”)
 名詞      動詞     助詞     動詞     助動詞    助動詞    助詞
"写真"    "撮っ"    "て"    "くれ"    "ませ"      "ん"      "か"
```

```
res[names(res) == "動詞"]
 動詞            動詞
"撮っ"          "くれ"
```

RMeCab パッケージによる解析

```
res <- .Call ("myMecab",” 写真撮ってくれませんか。”)
 名詞      動詞     助詞     動詞     助動詞    助動詞    助詞
"写真"    "撮っ"    "て"    "くれ"    "ませ"      "ん"      "か"
```

```
res[names(res) == "動詞"]
 動詞            動詞
"撮っ"          "くれ"
```

```
length(res)
      7
```

” 写真撮ってくれませんか” は 7 語 (形態素) からなる .

質問の文の長さを分析する

```
wilcox.test(Q3length, Q4length , paired = TRUE)
V = 1239, p-value = 5.327e-08
```

質問の文の長さを分析する

```
wilcox.test(Q3length, Q4length , paired = TRUE)
V = 1239, p-value = 5.327e-08
```

一般化線形モデル (GLM) による解析

- 目的変数: 質問文の長さ
- 説明変数 1: 出身地 (EW)
- 説明変数 2: 性別 (Sex)
- 説明変数 3: 遠慮の程度 (Q3A3, Q4A3)

質問の文の長さを分析する

```
wilcox.test(Q3length, Q4length , paired = TRUE)
V = 1239, p-value = 5.327e-08
```

一般化線形モデル (GLM) による解析

- 目的変数: 質問文の長さ
- 説明変数 1: 出身地 (EW)
- 説明変数 2: 性別 (Sex)
- 説明変数 3: 遠慮の程度 (Q3A3, Q4A3)

水準に順序 を導入した場合 (としない場合も 解析)

```
QA <- ordered (Q3A3, labels = c("L5", "L4", "H3", "H2", "H1"))
Levels: L5 < L4 < H3 < H2 < H1
```

GLM による文長の解析

158 名の回答を “A”, “B” の二つにわけ，“A” グループからは問 3(2)-(3) の解答を，“B” グループからは問 4(2)-(3) を抽出して統合

GLM による文長の解析

158 名の回答を “A”, “B” の二つにわけ，“A” グループからは問 3(2)-(3) の解答を，“B” グループからは問 4(2)-(3) を抽出して統合

- 分布族 : ポアソン分布 ， 負の二項分布
-

```
glm(Length ~EW + Sex + QA , family = "poisson",data = sample12)
```

GLM による文長の解析

158 名の回答を “A”, “B” の二つにわけ，“A” グループからは問 3(2)-(3) の解答を，“B” グループからは問 4(2)-(3) を抽出して統合

- 分布族 : ポアソン分布 ， 負の二項分布
-

```
glm(Length ~EW + Sex + QA , family = "poisson",data = sample12)
```

glm(Length ~EW + Sex + QA + pers (相手は大人か子供か), family = "poisson")

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.29635	0.09615	23.882	<2e-16 ***
EWW	0.02993	0.06253	0.479	0.6322
SexF	0.01666	0.06493	0.257	0.7975
QA.L	0.01289	0.20494	0.063	0.9498
QA.Q	0.05497	0.17558	0.313	0.7542
QA.C	0.11867	0.11724	1.012	0.3114
QÂ4	0.08216	0.06435	1.277	0.2017
persC	-0.10455	0.05228	-2.000	0.0455 *

Null deviance: 130.15 on 158 degrees of freedom

Residual deviance: 121.98 on 151 degrees of freedom

抽出された文末の形態素

levels(gobi)

"ですか"	"ますか"	"(ませ) んか"	"なんで"	"(いい) すかー"
"ていい"	"(せぇ) へんから "	"かなあ"	"くれん"	"かな"
"します"	"(いい)っすか"	"(いい) かなー"	"もいい"	"て下さい"
"けど頼める "	"いいか"	"てください"	"けどいい"	"ですが"
"お願いできる "	"ですけど "	"くれない"	"(ませ) んかぁ "	"(しょ) うか"
"ございます"	"もらえへん"	"もらえます"	"くれへん"	"かなぁ "

抽出された文末の形態素

levels(gobi)

"ですか"	"ますか"	"(ませ) んか"	"なんで"	"(いい) すかー"
"ていい"	"(せぇ) へんから "	"かなあ"	"く れん"	"かな"
"しま す"	"(いい)っすか"	"(いい) かなー"	"もいい"	"て下さい"
"けど 頼める "	"いいか"	"てく ださい"	"けどいい"	"ですが"
"お願いできる "	"ですけど "	"く れない"	"(ませ) んかぁ "	"(しょ) う か"
"ございます"	"もらえへん"	"もらえます"	"く れへん"	"かなぁ "

- 単語レ ベルのカテゴリ 調整
  - てく ださい，て下さい

抽出された文末の形態素

levels(gobi)				
"ですか"	"ますか"	"(ませ) んか"	"なんで"	"(いい) すかー"
"ていい"	"(せぇ) へんから"	"かなあ"	"くれん"	"かな"
"します"	"(いい)っすか"	"(いい) かなー"	"もいい"	"て下さい"
"けど頼める"	"いいか"	"てください"	"けどいい"	"ですが"
"お願いできる"	"ですけど"	"くれない"	"(ませ) んかぁ"	"(しょ) うか"
"ございます"	"もらえへん"	"もらえます"	"くれへん"	"かなぁ"

- 単語レベルのカテゴリ 調整
  - てください，て下さい
- 構文のレベルのカテゴリ 調整
  - 欲しいと 思う，欲しいと 思わない
  - 南瓜 で解析可能

形態素の統合

- て下さいに統一
  - 写真撮って下さい
  - 写真撮ってください

形態素の統合

- て下さいに統一
  - 写真撮って下さい
  - 写真撮ってください
- ですがに統一
  - 写真をお願いしたいのですが
  - 撮ってもらいたいんですけど

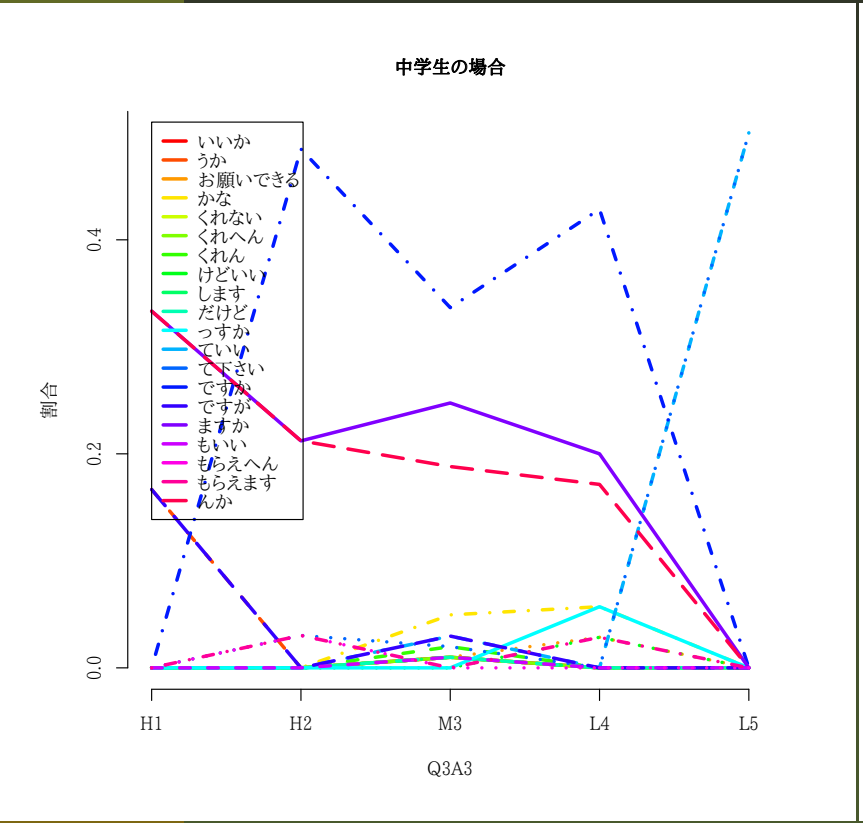
形態素の統合

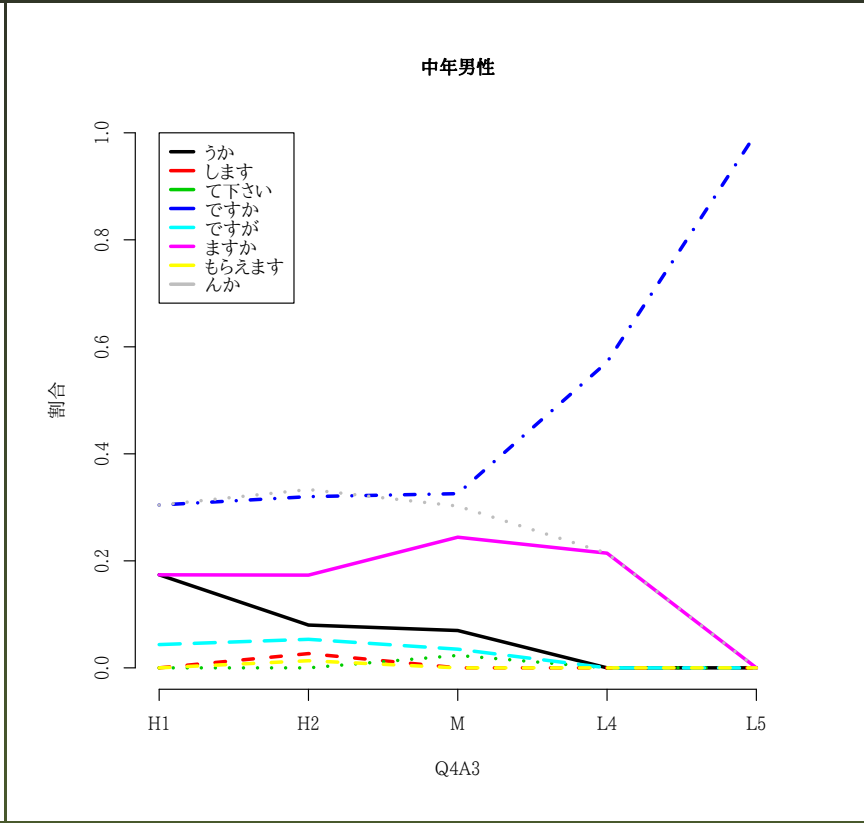
- て下さいに統一
  - 写真撮って下さい
  - 写真撮ってください
- ですがに統一
  - 写真をお願いしたいのですが
  - 撮ってもらいたいんですけど

```
data[gobi == "かなー", "gobi"] <- "かな" # 併合の処理
data[gobi == "かなあ", "gobi"] <- "かな"
data[gobi == "かなぁ", "gobi"] <- "かな"
levels(gobi)
```

"ですか"	"ますか"	"(ませ)んか"	"(いい)っすか"	"ていい"
"だけど"	"かな"	"くれん"	"します"	"もいい"
"て下さい"	"お願いできる"	"いいか"	"けどいい"	"ですが"
"くれない"	"(しょ)うか"	"もらえへん"	"もらえます"	"くれへん"

語尾の利用率

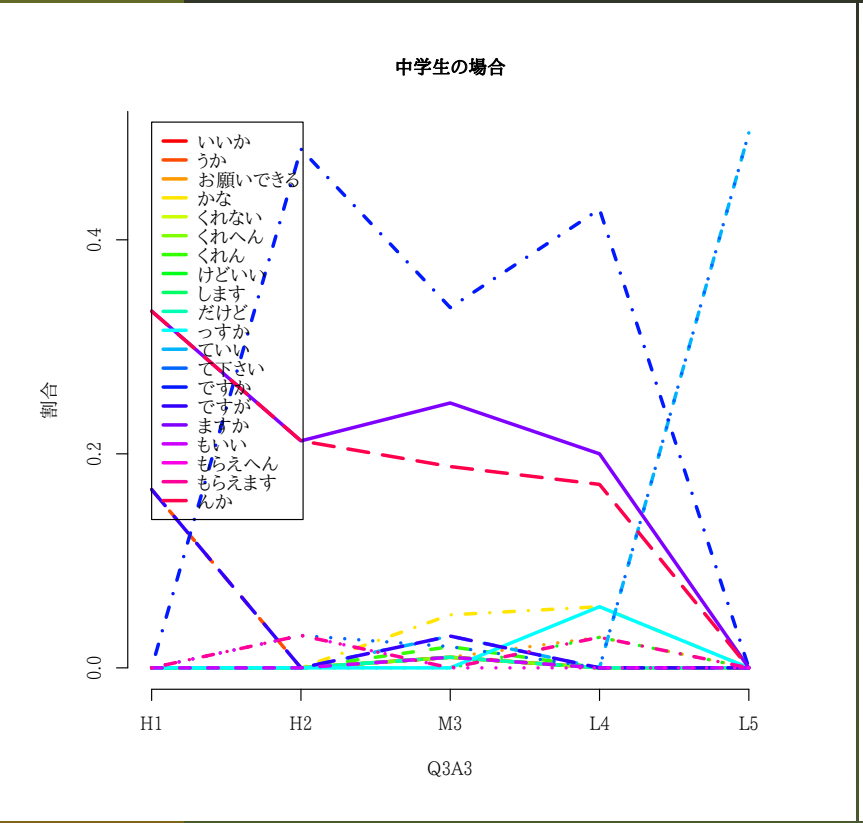


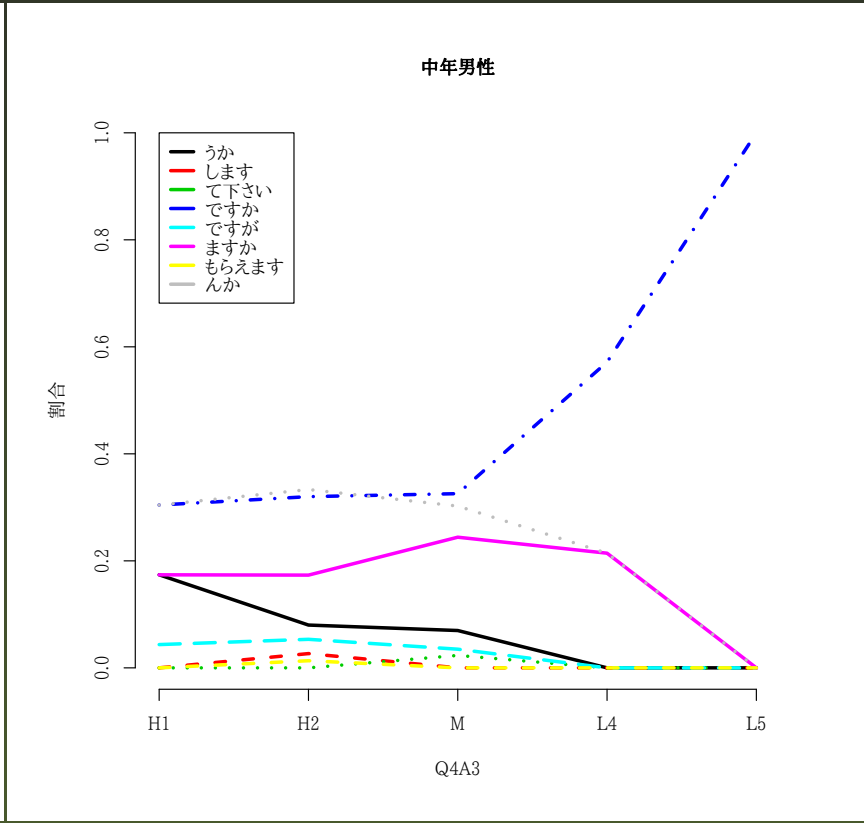


横軸:H1(「 非常に気をつかう 」),H2(かなり 気をつかう),M1(少し は気をつかわない),L4(あまり 気をつかわない),L5(まったく 気をつかわない)

縦軸:各水準ごと にみた語尾の割合

語尾の利用率





横軸:H1(「 非常に気をつかう 」),H2(かなり 気をつかう),M1(少し は気をつかわない),L4(あまり 気をつかわない),L5(まったく 気をつかわない)

縦軸:各水準ごと にみた語尾の割合

中学生: (ませ) んか，(でしょ) う か -> 減少へ

中年男性: (ませ) んか，ますか -> 減少へ

対応分析によるグラフ

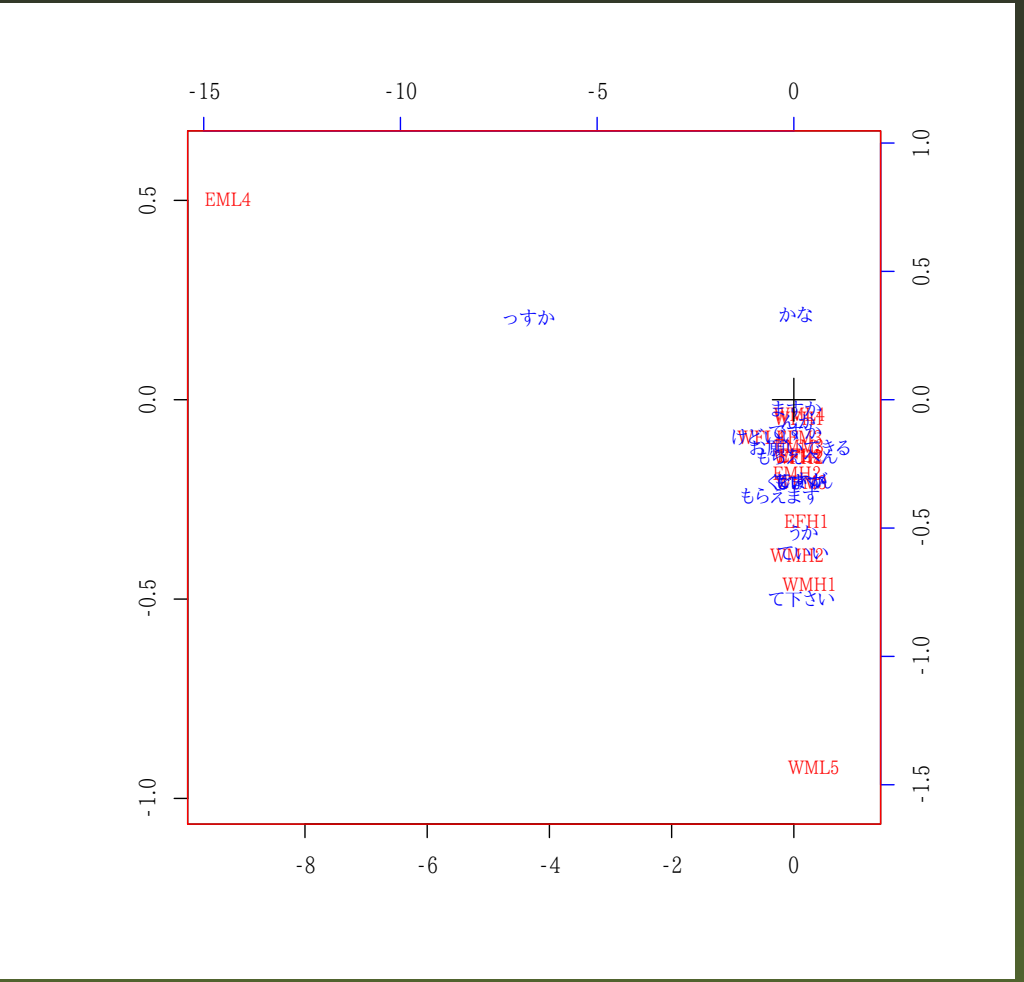
相手が中学生の場合

```
dat.t1 <- ftable(xtabs(~EW
+ Sex + Q3A3 + gobi,
    data = dat))
dat.t2 <- dat.t1 [row-
Sums(dat.t1) != 0, ]
dat.corr <- corresp (dat.t2,
      nf = min(nrow(dat.t2),
ncol(dat.t2) ))
biplot(dat.corr)
```

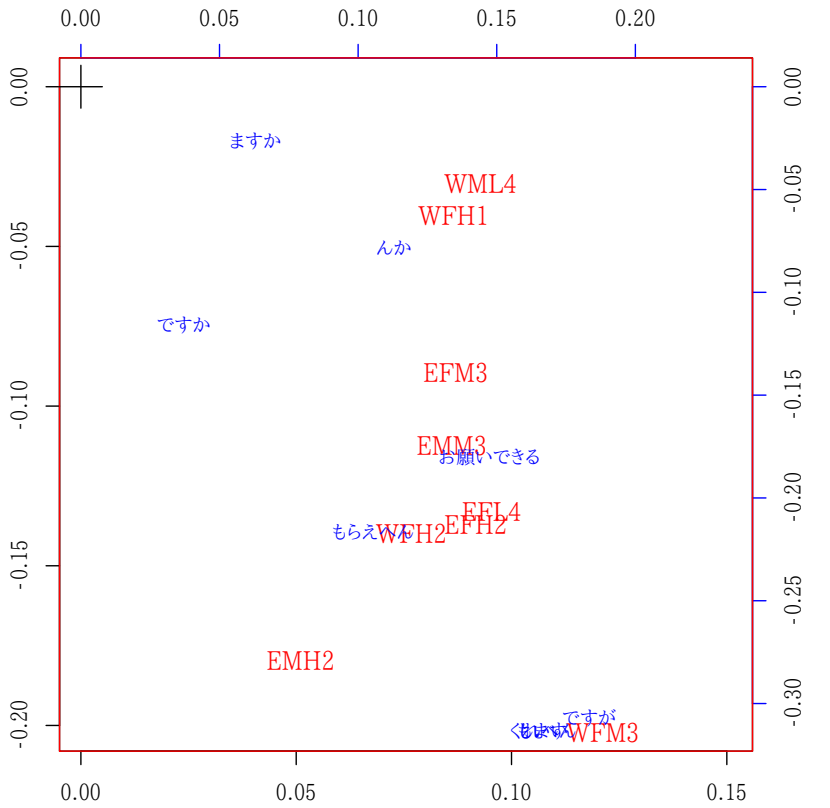
対応分析によるグラフ

相手が中学生の場合

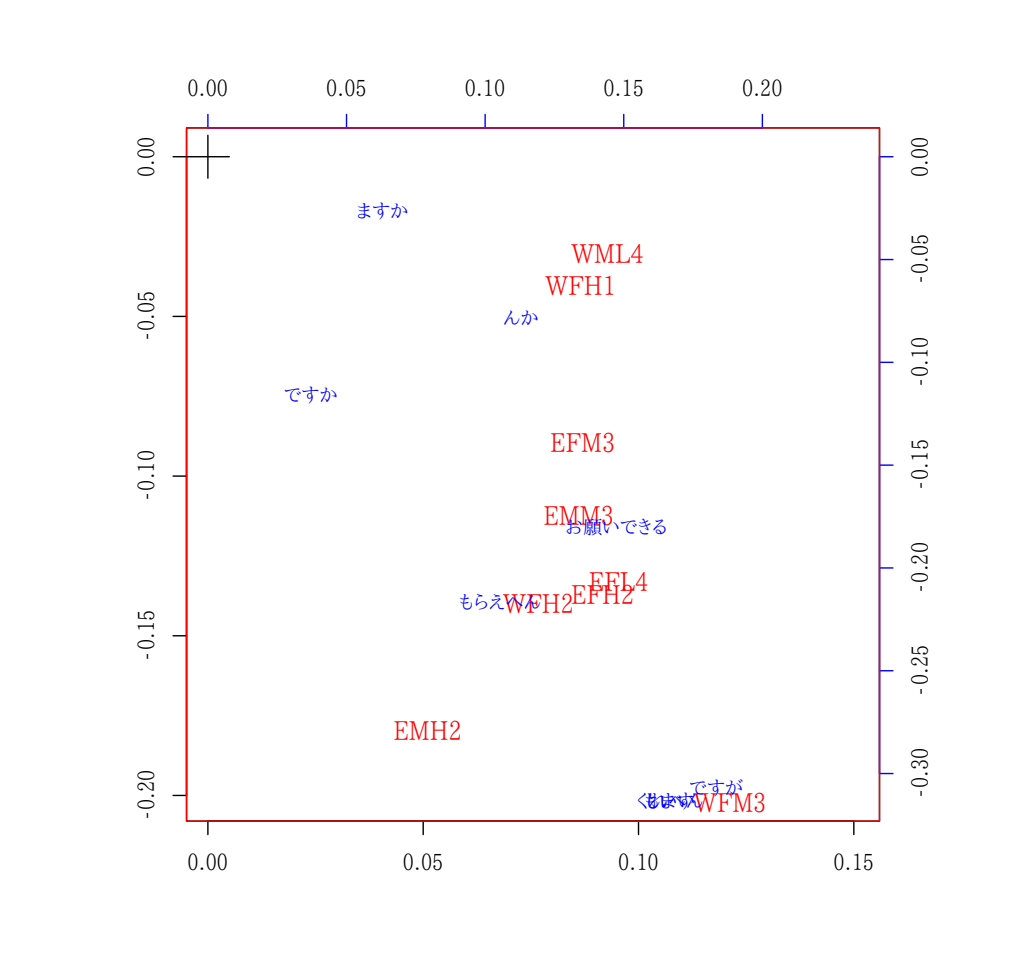
```
dat.t1 <- ftable(xtabs(~EW
+ Sex + Q3A3 + gobi,
    data = dat))
dat.t2 <- dat.t1 [row-
Sums(dat.t1) != 0, ]
dat.corr <- corresp (dat.t2,
      nf = min(nrow(dat.t2),
ncol(dat.t2) ))
biplot(dat.corr)
```



対人遠慮と 出身地域 1


0.00
0.05
0.10
0.15
0.20
-0.00
-0.05
-0.10
-0.15
-0.20
-0.25
-0.30
ますか
WML4
WFH1
んか
ですか
EFM3
EMM3
お願いできる
EFL4
EFH2
WFH2
EMH2
ですが
WFM3

対人遠慮と 出身地域 1



- お願いできる？
- もらえへん？

EMM3

(東日本:男性:遠慮度 中)

EFM3

(東日本:女性:遠慮度 中)

WFH2

(西日本:女性:遠慮度 やや高い)

EFH2

(東日本:女性:遠慮度 やや高い)

EFL4

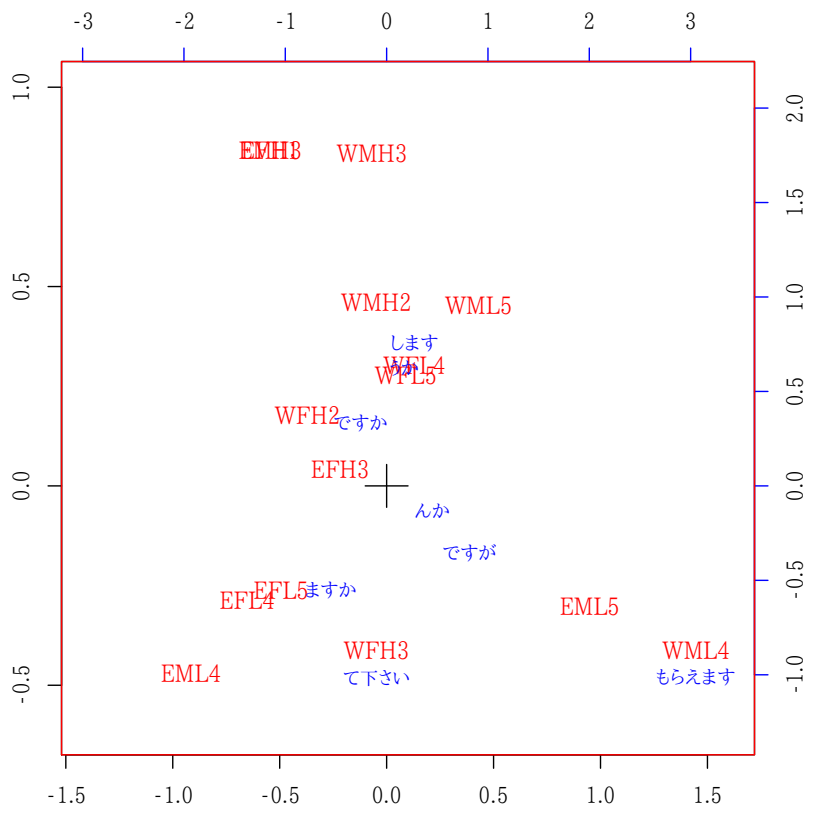
(東日本:女性:遠慮度 やや低い)

対人遠慮と 出身地域2

相手が中年男性の場合

"ですか"　"ますか"　"んか"　"ですが"

"うか"　"て下さい"　"します"　"もらえます"

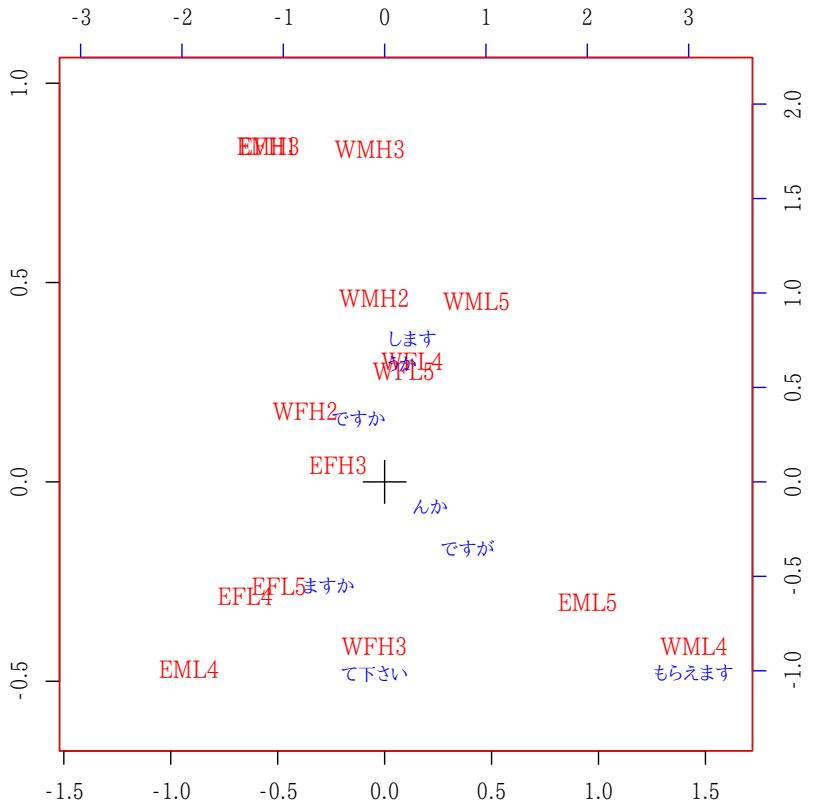


対人遠慮と 出身地域2

相手が中年男性の場合

"ですか"　"ますか"　"んか"　"ですが"

"うか"　"て下さい"　"します"　"もらえます"



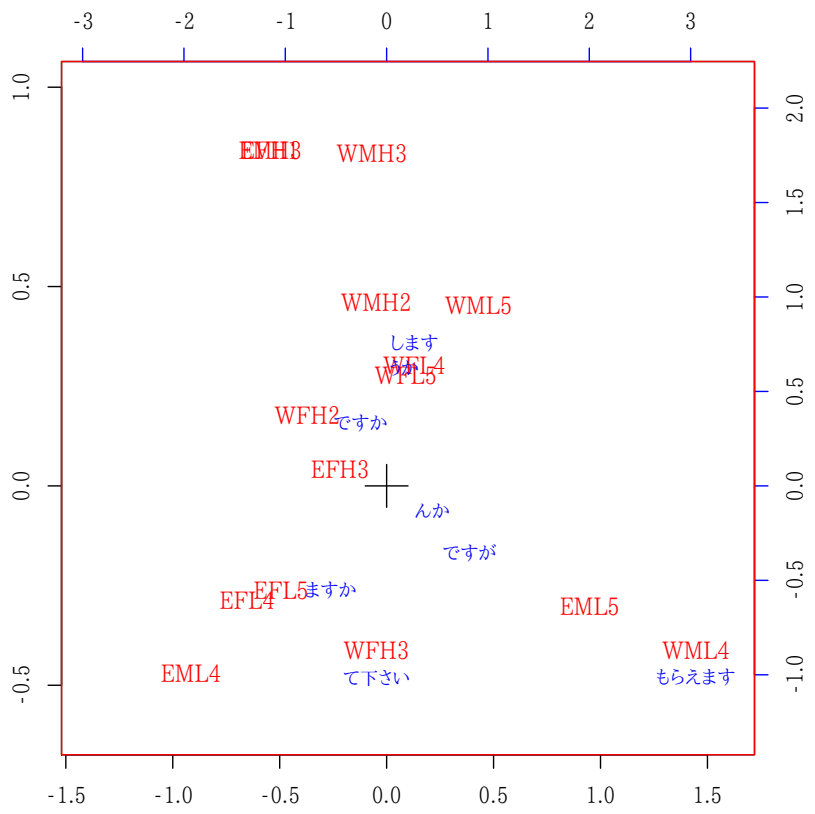

```
# 語尾の分布に多項ロジット 分析
library(nnet)
model1 <- multinom(gobi ~EW + Sex + Q4A3)
model2 <- step (model1)
summary(model2)
# 最終モデル
# 語尾のカテゴリ を出身地で説明する モデル
multinom(formula = gobi ~EW)
```

対人遠慮と 出身地域2

相手が中年男性の場合

"ですか"　"ますか"　"んか"　"ですが"

"うか"　"て下さい"　"します"　"もらえます"




```
# 語尾の分布に多項ロジット 分析
library(nnet)
model1 <- multinom(gobi ~EW + Sex + Q4A3)
model2 <- step (model1)
summary(model2)
# 最終モデル
# 語尾のカテゴリ を出身地で説明する モデル
multinom(formula = gobi ~EW)
```

```
#「 (ませ) んか」の有無をロジット 分析
glm(gobi ~EW + Sex + Q4A3, family = bino-
mial)
```

	Estimate	z value	Pr(>|z|)
EWW	0.881919	2.055	0.0399 *

おわり

- ご清聴ありがとうございました.
- 参考資料など
  - A Language and Environment for Statistical Computing: **R**,http://cran.r-project.org/
  - Taku Kudo : **MeCab**, http://mecab.sourceforge.net/
  - M. Konchady: Text Mining Application,2006
  - J.J. Faraway: Extending the Linear Model with R, 2005
  - Manning & Schuetze: Foundations of Statistical Natural Language Processing, 1999
  - Lebart & Salem & Berry: Exploring Textual Data, 1998