

Lusitanistentag 2009

Vergleich statistischer Methoden in der Korpuslinguistik am Beispiel einer soziolinguistischen Untersuchung des gesprochenen Portugiesisch

Peter Bouda, p.bouda@gmx.de

Gliederung

- Die Fragestellung
- Das Korpus
- Über statistische Methoden
- χ^2 -Test
- Mann-Whitney-Test
- Vergleich und Folgerungen

Die Fragestellung

- Unterscheidet sich der Sprachgebrauch bei männlichen und weiblichen Sprechern?
- Genauer: Inwieweit benutzen Männer und Frauen unterschiedliche Wörter?
- vergl.: Rayson et al. (1997) für das Englische
- Beschränkung auf gesprochene Sprache
- Untersuchung anhand eines Korpus

Das Korpus

- erster Teil: Ferreira (unveröffentlicht), "The Construction of Portuguese"
- gesammelt von April 2002 bis Oktober 2002
- Insgesamt 42 Sprecher: 17 männlich, 25 weiblich

	gesamt	mask	fem
Tokens	177003	61371	115632
Types	10294	5537	7760

Das Korpus

- Zweiter Teil: Português Falado - Variedades Geográficas e Sociais (Bacelar do nascimento 2001), Portugal Anos 90
- gesammelt von 1995-1997
- Insgesamt 20 Sprecher: 10 männlich, 10 weiblich

	gesamt	mask	fem
Tokens	19183	9970	9213
Types	3088	2123	1676

Das Korpus

- Insgesamt also 62 Sprecher: 35 weiblich, 27 männlich

	gesamt	mask	fem
Tokens	196186	71341	124845
Types	11381	6483	8273

Das Korpus

- Mittelwerte und Variationskoeffizienten (=Standardabweichung/Mittelwert), pro Text:

	mask_mean	mask_var	fem_mean	fem_var
Tokens	2642.26	0.709	3567	0.721
Types	642.56	0.487	746.66	0.521

Signifikanztests

- Nullhypothese H_0 vs. Alternativhypothese H_1
- **H_0 : Das Wort "casa" wird in unserem Korpus von Frauen und Männern gleich häufig verwendet.**
- **H_1 : Das Wort "casa" wird in unserem Korpus von Frauen oder Männern häufiger verwendet**
- Statistik berechnet Wahrscheinlichkeit dafür, dass H_0 gilt

Signifikanztests

- Die Alternativhypothese kann als bewiesen angenommen werden, wenn die Wahrscheinlichkeit p für die Nullhypothese einen Schwellenwert unterschreitet
- z.B. bei $p < 0,05$ bzw. konservativer $p < 0,01$
- "zweiseitige" vs. "einseitige" Hypothesen, hat u.U. Auswirkung auf den Test

Die Statistik - χ^2

- Wie oft kommt "casa" in den beiden Korpusanteilen vor?

	mask	fem	gesamt
"casa"	118	437	555
andere	71223	124408	195631
gesamt	71341	124845	196186

$$\chi^2 = \sum \frac{(\text{beobachtet} - \text{erwartet})^2}{\text{erwartet}}$$

$$\text{erwartet} = \frac{\text{gesamt_in_Zeile} * \text{gesamt_in_Reihe}}{\text{gesamt_gesamt}}$$

Die Statistik - χ^2

- Erwartete Werte und χ^2 -Test:

	mask	fem	gesamt
"casa"	118 (201,82)	437 (353,18)	555
andere	71223 (71139,18)	124408 (124491,82)	195631
gesamt	71341	124845	196186

- $\chi^2 = 54,86$, $df=1$, $p=1,294e-13$
- Also: Nullhypothese widerlegt (für $p<0,01$)
- "casa" wird signifikant häufiger von Frauen als von Männern verwendet

χ^2 -Test

- **Vorsicht 1: Der χ^2 -Test sagt nichts über die "Stärke" des Effekts aus**
- dafür gibt es den phi-Wert bzw. Cramer's V (Gries 2005, Gries 2009)
- normiert zwischen 0 (sehr schwacher Effekt) und 1 (sehr starker Effekt)
- für "casa": $\Phi = 0,0167$
- Die Korrelation zwischen "weiblich" und dem Gebrauch von "casa" ist sehr schwach

χ^2 -Test

- **Vorsicht 2: Neben der Verteilung zwischen den (beiden) Korpusteilen spielt auch die Verteilung innerhalb der Teile ein Rolle**
- Wurde das Wort nur von einem/zwei/... Sprecher häufig verwendet?
- dafür gibt es Verteilungstest wie z.B. Juilland et. al.'s D oder Gries' DP (Gries 2008)
- für "casa": DP = 0,23722, D = 0,89838
- relativ homogene Verteilung

χ^2 -Test

- **Vorsicht 3: Sprache und Sprachdaten entsprechen nicht dem "Zufallsmodell", das normalerweise der Statistik zugrunde liegt**
- "Library Metaphor" von Evert (2006)
- vorgeschlagen z.B.: Student's t-test (Evert 2009)
- bisher keine allgemeine Lösung
- "pro" Signifikanztests: binomiale Verteilung als untere Grenze für Variation in Sprache (Evert 2009)

χ^2 -Test

- **Vorsicht 4: Je mehr Beweise man hat, desto signifikanter sind die Ergebnisse**
- Kilgarriff (2001): mit einem Korpus, der groß genug ist, kann man alles beweisen
- Signifikanztests sind quantitativ richtig, aber qualitativ falsch (Gries 2005)
- Kilgarriff schlägt als Alternative den Mann-Whitney-/Wilcox-Test vor

χ^2 -Test

- Zwischenergebnis "Sprache der Frauen":

Rang	Wort	f mask	f fem	χ^2	p	Φ	D	DP
1	eu	981	2549	114,1857	0,0000	0,0241	0,9374	0,1595
2	ele	204	732	86,2667	0,0000	0,0210	0,8452	0,3565
3	casa	118	437	54,8603	0,0000	0,0167	0,8984	0,2372
4	meu	228	685	51,4366	0,0000	0,0162	0,8942	0,2383
5	lhe	100	377	49,0027	0,0000	0,0158	0,8639	0,3395

Mann-Whitney-Test

- der Test nicht-parametrisch, d.h. die Korpusgröße hat keine Auswirkung auf die Stärke der Korrelation
- Mann-Whitney-Test berücksichtigt Ein-/Zweiseitigkeit der Hypothese
- hier: zweiseitige Hypothese
- alternativer Signifikanztest für den Korpusvergleich (Kilgarriff 2001)

Mann-Whitney-Test

- Beide Korpusteile werden in mehrere, gleich große Teile geteilt (mask: 31; fem: 54)
- Diese Teile werden zufällig zu größeren Einheiten zusammengesetzt (mask: 5; fem: 9)
- In jedem Teil wird das Wort "casa" gezählt
- Die Teile werden in der resultierenden Reihenfolge aufgelistet

Mann-Whitney-Test

- Reihenfolge für "casa":

Anzahl	17	21	22	24	30	41	43	45	47	48	49	51	53	60	s
Korpus	m	m	m	m	m	f	f	f	f	f	f	f	f	f	
Rang (f)						6	7	8	9	10	11	12	13	14	90
Rang (m)	1	2	3	4	5										15

- es wird ein Signifikanztest über die Rangsumme durchgeführt
- für "casa": $p = 0,0001$
- Nullhypothese verworfen bei $p < 0,01$

Mann-Whitney-Test

- Vorteile:
 - besonders hohe Frequenzen in einzelnen Teilen haben keinen Einfluss auf p
 - p ist nicht abhängig von der Korpusgröße
- Nachteile:
 - p ist nicht abhängig von der Korpusgröße [sic]
 - durch die Rangsumme wird eine zusätzliche "Schicht" in die Interpretation der Daten eingeführt
 - der Test ist nicht besonders gut erforscht

Vergleich

- χ^2 vs. Mann-Whitney (Top 10):

Rang	Wort	χ^2	Korp
1	eu	114,19	fem
2	ele	86,27	fem
3	pá	73,50	mask
4	la	69,31	mask
5	tropa	62,17	mask
6	casa	54,86	fem
7	futebol	52,49	mask
8	meu	51,44	fem
9	lhe	49,00	fem
10	mãe	45,91	fem

Rang	Wort	p	Korp
1	eu	0,00243	fem
2	casa	0,00382	fem
3	houve	0,00552	mask
4	oito	0,00793	mask
5	pára	0,00811	mask
6	ai	0,00871	fem
7	tropa	0,00988	mask
8	marido	0,01005	fem
9	sozinha	0,01013	fem
10	mãe	0,01030	fem

Folgerungen I

- der Korpuslinguist hat die Qual der Wahl bei den Tests
- statistische Tests bereichern die Korpuslinguistik um anerkannte Methoden
- Signifikanztests sind wichtig!
- aber: mindestens eine zusätzliche Interpretationsschicht

Folgerungen II

- Tests sollten möglichst einfach und intuitiv verständlich sein
- aktuelle Debatte: Wie "zufällig" sind Sprachdaten? (Evert 2006 und Evert 2009)
- hier fehlt Forschung über statistische Methoden speziell für Korpuslinguistik
- Sprache ist immer komplexer als jede Statistik!
- Loftus (1991): "On the Tyranny of Hypothesis Testing in the Social Sciences"

Ergebnisse I

- Top 10 "Sprache der Männer"

RANG	WORT	F_MASC	F_FEM	CHI	PHI	JUILLAND_D	GRIES_DP
1	pá	161	99	73,49650	0,01936	0,54151	0,75761
2	la	73	21	69,30579	0,01880	0,30281	0,74893
3	tropa	56	12	62,17170	0,01780	0,70350	0,69030
4	futebol	34	2	52,48900	0,01636	0,16768	0,84721
5	eh	142	109	44,35852	0,01504	0,74319	0,72099
6	linho	25	0	43,75496	0,01493	-0,00816	0,99569
7	sim	410	472	39,22146	0,01414	0,83319	0,40475
8	bocado	92	61	37,37626	0,01380	0,79116	0,52219
9	maneira	64	37	31,84012	0,01274	0,74386	0,49809
10	ora	48	22	31,39004	0,01265	0,80095	0,53600

Ergebnisse II

- Top 10 "Sprache der Frauen"

RANG	WORT	F_MASC	F_FEM	CHI	PHI	JUILLAND_D	GRIES_DP
1	eu	981	2549	114,1857	0,0241	0,93739	0,15947
2	ele	204	732	86,2667	0,0210	0,84516	0,35646
3	casa	118	437	54,8603	0,0167	0,89838	0,23722
4	meu	228	685	51,4366	0,0162	0,89419	0,23830
5	lhe	100	377	49,0027	0,0158	0,86394	0,33951
6	mãe	86	334	45,9126	0,0153	0,87157	0,32823
7	ela	158	502	44,1769	0,0150	0,87140	0,31277
8	ai	26	163	41,7835	0,0146	0,83877	0,41593
9	assim	444	1099	38,7055	0,0140	0,91553	0,20937
10	minha	282	739	33,9106	0,0131	0,88524	0,28977

Schluss

Vielen Dank für Ihre Aufmerksamkeit!

Download Präsentation: www.peterbouda.de

Literatur

- Bacelar do Nascimento, M. F. (coord.) (2001), 'Português Falado - Variedades Geográficas e Sociais', Centro de Linguística da Universidade de Lisboa e Instituto Camões, URL: www.clul.ul.pt/sectores/linguistica_de_corpus/projecto_portuguesfalado.php.
- Evert, S. (2009), 'Rethinking Corpus Frequencies', ICAME 2009 Handout, URL: www.cogsci.uni-osnabrueck.de/.../Evert2009_ICAME.handout.pdf.
- Evert, S. (2006), 'How Random is a Corpus? The Library Metaphor', *Zeitschrift für Anglistik und Amerikanistik* 54(2), 177-190.
- Ferreira, V. (unveröffentlicht), 'The Construction of Portuguese', PhD thesis, Ludwig-Maximilian-University Munich.
- Gries, S. Th. (2009), *Quantitative Corpus Linguistics with R*, London/New York: Routledge.
- Gries, S. Th. (2008), 'Dispersions and adjusted frequencies in corpora', *International Journal of Corpus Linguistics* 13:4, 403-437.
- Gries, S. Th. (2005), 'Null-hypothesis significance testing of word frequencies: a follow-up on Kilgarriff', *Corpus Linguistics and Linguistic Theory* 1-2, 277-294.
- Kilgarriff, A. (2001), 'Comparing Corpora', *International Journal of Corpus Linguistics* 6:1, 1-37.
- Loftus, G. R. (1991), 'On the Tyranny of Hypothesis Testing in the Social Sciences', *Contemporary Psychology* 2, 102-105.
- Paul Rayson, G. L. & Hodges, M. (1997), 'Social Differentiation in the Use of English Vocabulary: Some Analyses of the Conversational Component of the British National Corpus', *International Journal of Corpus Linguistics* 2(1), 133-152.