EURAC research

Evaluating the use of idioms in an L1 learner corpus

Andrea Abel

Aivars Glaznieks

Verena Blaschitz

Institute for Specialised Communication and Multilingualism, Bolzano/Bozen (Italy)

the phenomenon

to brush one's teeth vs. it. *lavarsi i denti* "to wash …"
dt. *sich die Zähne putzen* "to clean …"
…

varieties + formulaic sequences

differences appear on different levels (cf. Földes 1996):

- phonetic/prosodic: *etw. springt jmdm. ins Aug‘* (vs. *Auge*)
- word formation: *bis aufs I-Tüpferl* (vs. *Tüpfelchen*)
- morphosyntactic:
  - gender: *zerrinnen wie der* (vs. *die*) *Butter in der Sonne*
  - prepositions: *am* (vs. *auf dem*) *Prüfstand stehen*
  - valency: (vs. *einen*) *über den Durst trinken*
  - inflection: auxiliary *sein* (vs. *haben*) with *sitzen, stehen …*
- lexical: *einen Knödel* (vs. *Kloß*) *im Hals haben*
- semantic content: *deutsches Eck* ‘the shortest road and railway link between Salzburg and Innsbruck’ (vs. ‘headland in Koblenz‘)

overview

1. background information
    - formulaic language + (l1) learner
2. l1 learner corpus KoKo
    - the project behind
    - method of corpus analysis
    - results
    - interpretation
3. summary/conclusion

formulaic language + (l1) learner

important functions:

- support daily communication (“facilitation”)
- indicate the membership to a language group (“identification”)

in school:

- source of errors (cf. Margewitsch 2006, Eichler 2004)
- decreasing usage of phraseology (cf. Wodak & Rheindorf 2011)
- “noticeable” usage of phraseology (cf. Dürscheid et al. 2010)

at university level:

- collocations as indicator for competent academic writing (cf. Steinhoff 2007)

the project *KoKo*

full titel: comparing *Bildungssprache*: analysis of the language competence of - especially South Tyrolean - German L1 learners on the basis of corpora

partners: EURAC research; Freie Universität Bozen / Libera Università di Bolzano / Free University of Bozen · Bolzano; universität innsbruck

KoKo is part of the initiative *Korpus Südtirol*

http://www.korpus-suedtirol.it

Korpus Südtirol

the *KoKo* corpus

- *text type*: argumentative texts on a provocative statement about youth taken from an interview with the author H.M. Enzensberger
- *participants*: pupils from secondary schools, 1 year before their school-leaving examination (random sampling), age 17-19; from South Tyrol (Italy), North Tyrol (Austria), Thuringia (Germany)
- *metadata*: age, gender, region, L1/L2, school type, population of the location of the school, socio-economic background, linguistic biography (dialect vs. standard variety)

the *KoKo* corpus

Sub-corpus (region)	total		L1 German	
	tokens	texts	tokens	texts
North Tyrol:	233,098	457	206,439	404
South Tyrol:	222,209	520	192,891	451
Thuringia:	353,674	521	317,075	464
not defined for region	2349	5	---	---
total	811330	1503	716,405	1319

automatic annotation for lemma an POS using the TreeTagger (Schmidt 1994)

observations

„Der Aussage des Deutschen kann ich nicht ganz zu stimmen, da er alle Jugendlichen unter einen Hut steckt." (ID1762)

I cannot agree with the claims of the German, since he lumps together all the young people (lit. to put all the young people under one hat)." (ID1762)

etw. unter einen Hut bringen 'to juggle/reconcile sth.'

alle in einen Topf werfen 'to lump together'

alle unter einen Hut stecken 'to lump together'

research focus

general research question:

how can learner errors be separated from other forms of variation (esp. diatopic, diachronic)?

focus of analysis: use of idioms

method

corpus linguistic approaches:

- automatic extraction following a given pattern of formulaic sequences (cf. Weller & Heid 2010)
- query of specific formulaic sequences (cf. Fellbaum 2007)
- combined approaches (cf. Quasthoff et al. 2010)

problems:

1. low frequency: idioms, proverbs, and similes are found less frequent than collocations such as discourse markers/ connectives (cf. Moon 2007: 1049-1051)
2. corpus size: “a large corpus is needed to find natural occurrences of idiomatic expressions” (Sailer 2007: 1062)

recursive approach

- combination of queries following abstract pattern (*consultation*) and specific sequences (*analysis*)

step 1: collocation extraction (cf. Weller & Heid 2010):

query for *APPR (+ DET) (+ ADJA) + NN + VV*

step 2: analysis of candidates for idioms

step 3: sorting for key words (NN)

step 4: corpus query considering key words (NN)

step 5: verification of the idiom

step 1: collocation extraction

query for APPR (+ DET) (+ ADJA) + NN + VV: 27.122 hits

Frequency count for lemma (APPR NN VV) (*N*(types)=18.469):

95	in	indef	-	Interview	sagen	Sg
85	aus	-	-	Fehler	lernen	Pl
57	in	-	voll	Zug	genießen	Pl
49	in	indef	-	Interview	äußern	Sg
48	mit	-	-	Freund	treffen	Pl
45	in	def	-	Diskothek	gehen	Sg
43	zu	-	-	Schule	gehen	Sg
43	in	-	-	Kopf	haben	Sg
43	aus	poss	-	Fehler	lernen	Pl
40	als	-	-	Jugendlich	haben	PlSg
37	in	indef	-	Topf	werfen	Sg
35	in	def	-	Jugend	machen	Sg
34	in	def	-	Jugend	haben	Sg
(…)						

step 2: analysis

query for APPR (+ DET) (+ ADJA) + NN + VV: 27.122 hits

classifying candidates for idioms (i.S.v. Burger 2007: 63)

95	in	indef	-	Interview	sagen	Sg
85	aus	-	-	Fehler	lernen	Pl
57	in	-	voll	Zug	genießen	Pl
49	in	indef	-	Interview	äußern	Sg
48	mit	-	-	Freund	treffen	Pl
45	in	def	-	Diskothek	gehen	Sg
43	zu	-	-	Schule	gehen	Sg
43	in	-	-	Kopf	haben	Sg
43	aus	poss	-	Fehler	lernen	Pl
40	als	-	-	Jugendlich	haben	PlSg
37	in	indef	-	Topf	werfen	Sg
35	in	def	-	Jugend	machen	Sg
34	in	def	-	Jugend	haben	Sg
(…)						

step 3: sorting for key words

example: all idioms with NN = Topf (‘pot’), filter: L1=German

32	in	indef	-	Topf	werfen	Sg
10	in	indef	-	Topf	stecken	Sg
5	in	indef	-	Topf	schmeißen	Sg
2	in	indef	-	Topf	hauen	Sg
2	in	dem	-	Topf	werfen	Sg
2	in	def	gleich	Topf	schmeißen	Sg
1	in	dem	-	Topf	stecken	Sg
1	in	dem	-	Topf	schmeißen	Sg
1	in	def	gleich	Topf	werfen	Sg
1	in	def	gleich	Topf	stecken	Sg
57						

step 4: corpus query

example: all idioms with NN = Topf ('pot'), filter: L1=German

38	in	indef	-	Topf	werfen	Sg
10	in	indef	-	Topf	stecken	Sg
7	in	indef	-	Topf	schmeißen	Sg
2	in	indef	-	Topf	hauen	Sg
4	in	dem	-	Topf	werfen	Sg
2	in	def	gleich	Topf	schmeißen	Sg
1	in	dem	-	Topf	stecken	Sg
1	in	dem	-	Topf	schmeißen	Sg
1	in	def	gleich	Topf	werfen	Sg
1	in	def	gleich	Topf	stecken	Sg
67						

step 5: verification

ID1041(NT): „Ein junger Mensch ist womöglich noch unsicher und leichter beeinflussbar, als eine Person, die schon etwas erfahrener ist, aber genau aus diesem Grund ist es wichtig, die Jugend nicht zu kritisieren, sondern ihnen zu helfen, die richtigen Entscheidungen zu treffen und sie auf den richtigen Weg zu bringen. Meiner Meinung nach ist es falsch <**alle Jugendlichen in einen Topf zu werfen**>. Menschen machen Fehler, egal wie alt sie sind und man sollte sich in seiner Jugend auch austoben können und nicht nach einem strengen Schema leben müssen. “

ID2658(NT): „Wie bereits erwähnt, ist laut Hans Magnus Enzensberger die Jugend keine beneidenswerte Phase, denn viele Leiden beispielsweise an einer Klamottensucht und wollen immer das haben, was andere auch haben, damit sie "cool" sind und dazu gehören. Jedoch stellt sich die Frage, kann man wirklich <**alle Jugendlichen in einen Topf stecken**>? Meiner Meinung nach, gibt es genug Jugendliche, die ihre Freizeit nicht nur mit Diskothekbesuche, Alkohol und Tabak verbringen.“

results

27.122 hits for *APPR (+ DET) (+ ADJA) + NN + VV*, among them diff. expressions with the meaning 'to lump together':

- *alle Jugendlichen über einen Kamm scheren* "to shear all the young people with one comb"
- *alle Jugendlichen in einen Topf werfen* "to throw all the young people into one pot"
- *alle Jugendlichen auf eine Stufe stellen* "to put all the young people on one step"
- *alle Jugendlichen in einen Sack geben* "to put all the young people into one bag"
- *alle Jugendlichen unter einen Hut stecken* "to put all the young people under one hat"
- *alle jungen Menschen in eine Schublade stecken* "to put all the young people into one drawer"
- *alle in dieselbe Kiste stecken* "to put all people into the same box"
- *alle Jugendlichen in eine Ecke schieben* "to push all the young people into one corner"
- *alle jungen Menschen auf einen Haufen kehren* "to sweep all the young people onto one heap"

results

		'to lump together'
	KoKo: filter L1 German	
(1)	in DET (ADJA) Topf + VV	67
(2)	über DET (ADJA) Kamm + VV	14
(3)	auf DET (ADJA) Stufe + VV	3
(4)	unter DET (ADJA) Hut + VV	12
(5)	in DET (ADJA) Sack + VV	2
(6)	in DET (ADJA) Schublade + VV	14
(7)-(10)	...	4
	AKKO ADV(Dir) + VV	117

interpretation

default hypothesis:

H0 = the linguistic deviation x in *KoKo* is a learner error

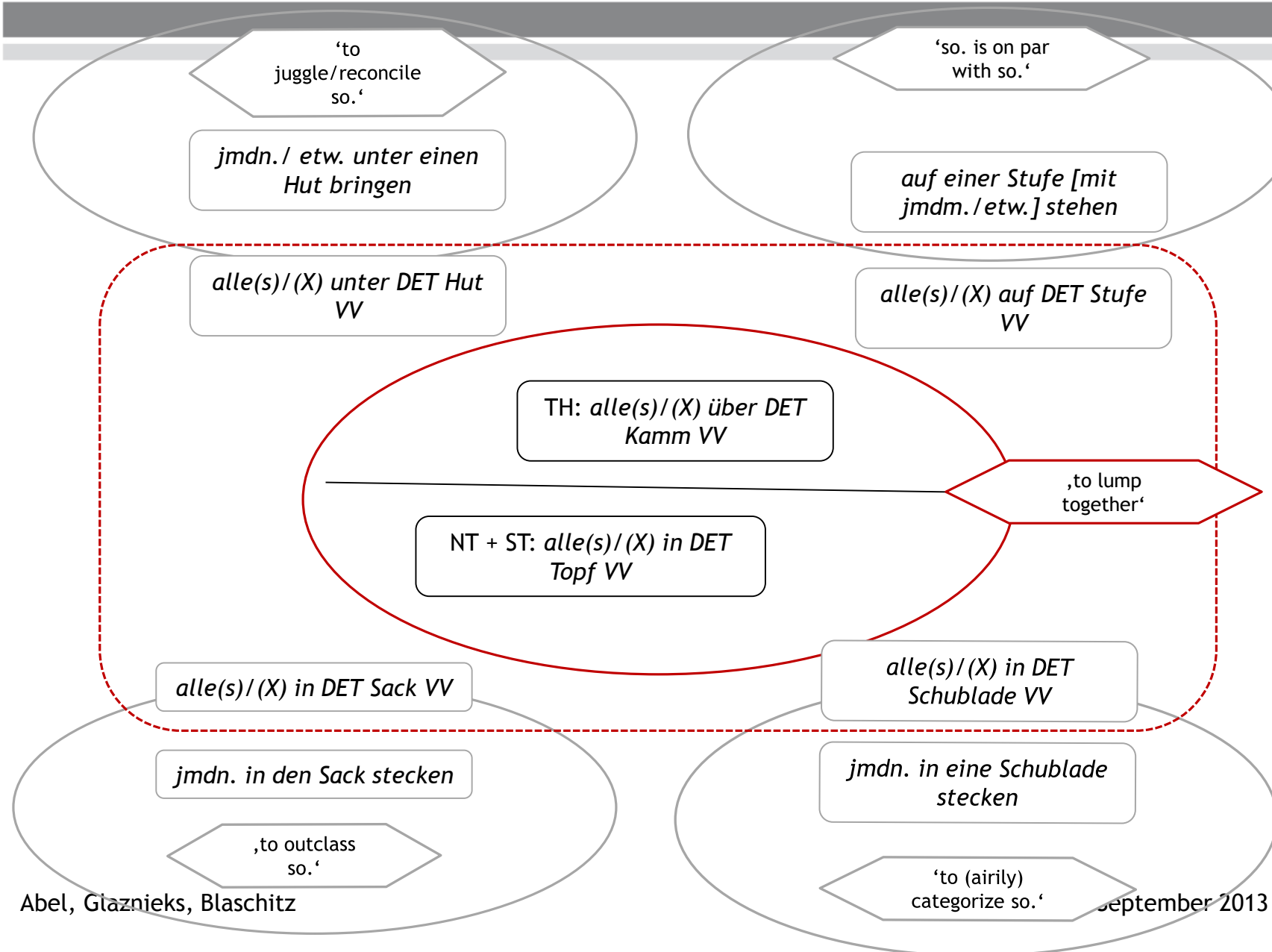
alternative hypotheses:

H1 = the linguistic deviation x in *KoKo* is a diatopic variant

H2 = the linguistic deviation x in *KoKo* is a phenomenon of language change

…

Hn = the linguistic deviation x in *KoKo* is a phenomenon of y


'to juggle/reconcile so.'
jmdn./ etw. unter einen Hut bringen
'so. is on par with so.'
auf einer Stufe [mit jmdm./etw.] stehen
alle(s)/(X) unter DET Hut VV
alle(s)/(X) auf DET Stufe VV
TH: alle(s)/(X) über DET Kamm VV
,to lump together'
NT + ST: alle(s)/(X) in DET Topf VV
alle(s)/(X) in DET Sack VV
alle(s)/(X) in DET Schublade VV
jmdn. in den Sack stecken
jmdn. in eine Schublade stecken
,to outclass so.'
'to (airily) categorize so.'

interpretation

alternative hypothesis:

H1 = the linguistic deviation x in *KoKo* is a diatopic variant

	'to lump together'				log likelihood		
KoKo: filter L1 German	total	NT	ST	TH	NT vs. ST	NT vs. TH	ST vs. TH
auf DET (ADJA) Stufe + VV	3	0	0	3	0.00†	-3,01	-2,85
unter DET (ADJA) Hut + VV	12	5	1	6	2,65	0,16	-1,91
in DET (ADJA) Sack + VV	2	0	2	0	-2.91†	0.00†	3.89†*
in DET (ADJA) Schublade + VV	14	6	4	4	0,28	1,72	0,49
AKKO ADV(Dir) + VV	113	44	41	28			

* = $p < 0.5$

† = frequency is too low for reliabal interpretation

interpretation

alternative hypothesis:

H2 = the linguistic deviation x in *KoKo* is a phenomenon of language change

	‚to lump together'			
	KoKo	AuNCo	DOL	GerNCo
	absolute	absolute	absolute	absolute
auf DET (ADJA) Stufe + VV	3 / 5	0 / 525	0 / 67	0 / 399
unter DET (ADJA) Hut + VV	12 / 16	3 / 1985	0 / 263	1 / 2002
in DET (ADJA) Sack + VV	2 / 2	4 / 20	0 / 2	12 / 24
in DET (ADJA) Schublade + VV	14 / 27	1 / 150	0 / 16	5 / 398

summary

candidates for learner errors:

- *alle Jugendlichen auf eine Stufe stellen* “to put all the young people on one step”
- *alle Jugendlichen unter einen Hut stecken* “to put all the young people under one hat”
- *alle jungen Menschen in eine Schublade stecken* “to put all the young people into one drawer”

candidate for language change:

alle Jugendlichen in einen Sack geben “to put all the young people into one bag”

conclusion

systematic procedure helps to avoid misinterpretation:

differentiating learner errors, diatopic variation, and phenomena of language change

outlook:

- consideration of other hypotheses for distribution
- extension of the comparison to reference corpora
- consideration of other types of formulaic sequences
- extension of the onomasiological analysis of the corpus

references

Duden (2008): *Redewendungen. Wörterbuch der deutschen Idiomatik*. Mannheim: Dudenverlag.

Dürscheid, C. et al. (2010): *Wie Jugendlich schreiben. Schreibkompetenz und neue Medien*. Berlin: de Gruyter.

Eichler, W. (2004): Sprachbewusstsein und grammatisches und stilistisches Formulieren: Falsche Kollokationen und verformelter Sprachgebrauch in Oberstufenaufsätzen. In: *ELiSe* 4 (1): 155-163.

Fellbaum, C. (2007) [Hg.]: *Idioms and Collocations*. London: Continuum.

Földes, C. (1996): *Deutsche Phraseologie kontrastiv*. Heidelberg: Groos.

Häcki Buhofer, A. (1998): Kenntnis- und Gebrauchsunterschiede bei Phraseologismen des Binnendeutschen, des Schweizerhochdeutschen und des Schweizerdeutschen. In: W. Eismann [Hg.]: *EUROPHRAS 1995*. S. 295-313. Bochum: Brockmeyer.

Margewitsch, E. (2006): *Formelhafter Sprachgebrauch in Schülertexten*. Oldenburg: Didaktisches Zentrum.

Moon, R. (2007): Corpus linguistic approaches with English corpora. In: H. Burger et al. [Hgg.]: *Phraseologie*. Vol. 2, S. 1045-1059. Berlin: de Gruyter. [= HSK 28.2]

Sailer, M. (2007): Corpus linguistic approaches with German corpora. In: H. Burger et al. [Hgg.]: *Phraseologie*. Vol. 2, S. 1060-1071. Berlin: de Gruyter. [= HSK 28.2]

Steinhoff, T. (2007): *Wissenschaftliche Textkompetenz*. Tübingen: Niemeyer.

Quasthoff, U. et al. (2010): Häufigkeit und Struktur von Phraseologismen am Beispiel verschiedener Web-Korpora. In: Ptashnyk et al. [Hgg.], S. 37-53.

Weller, M. & U. Heid (2010): Extraction of German multiword expressions from parsed corpora using context features. LREC 2010.

Wodak, R. & Rheindorf, M. (2011): *Wandel der deutschen Sprache: eine textsortenbezogene Pilotstudie*. Wien: Institut für Sprachwissenschaft.

Wardhough, R. (1998): *An Introduction to Sociolinguistics*. Blackwell.

Wray, A. & M. R. Perkins (2000): The functions of formulaic language: an integrated model. In: *Language & Communication* 20: 1-28.