Contrasts in English-German cohesion – frequencies, functional motivations, registerial variation and some impacts on translations

Erich Steiner
Saarland University
Saarbrücken

Acknowledgements

- Thanks to Marilisa Amoia, Kerstin Kunz, Ekaterina Lapshinova-Koltunski, and Katrin Menzel for co-operation in the GECCo- team

2. The necessity of predicting and operationalizing *across levels* and *between product and process*

2.1. Distinctive properties of translated text

2.2. Contrasts and contact on the level of cohesion

2.3. The relationship between processing and encoding

2.1. Distinctive properties of translated text

Definition of *explicitation* for Croco

> We assume explicitation if a translation (or language-internally one text in a pair of register-related texts) realizes meanings (not only ideational, but including interpersonal and textual) more explicitly than its source text – more precisely, meanings not realized in the less explicit source variant but implicitly present in a theoretically-motivated sense. The resulting text is more explicit than its counterpart.

(Hansen-Schirra, Neumann and Steiner 2012)

2.2. Contrasts and contact on the level of cohesion:

2.2.1. Types of phenomena

Cohesive reference and local underspecification

(5) And he answered them courteously that they should speak on, for he had not come so far and so wearily simply in order to turn back. Moreover he was charged by his father with a mission, which he might not reveal in that place. ‘It is known to us already,’ said the three damsels. [EO_FICTION_002]

(6) Und er erwiderte ihnen artig, daß sie weitersprechen sollten, denn er habe die Mühsal und Beschwerden des weiten Weges nicht auf sich genommen, um nun kehrtzumachen. Und zudem habe sein Vater ihn mit einer Aufgabe betraut, die er an diesem Ort zu enthüllen nicht gesonnen sei. ‘Dies ist uns bekannt’, sagten die drei Jungfrauen. [GTrans_FICTION_002]

Byatt, A.S. 1991. *Possession*. London: Vintage Books; Translation by Melanie Walz 1994 as *Besessen* Frankfurt/M.: Büchergilde Gutenberg

For discussion cf. Kunz and Steiner 2012

Local differences in cohesion

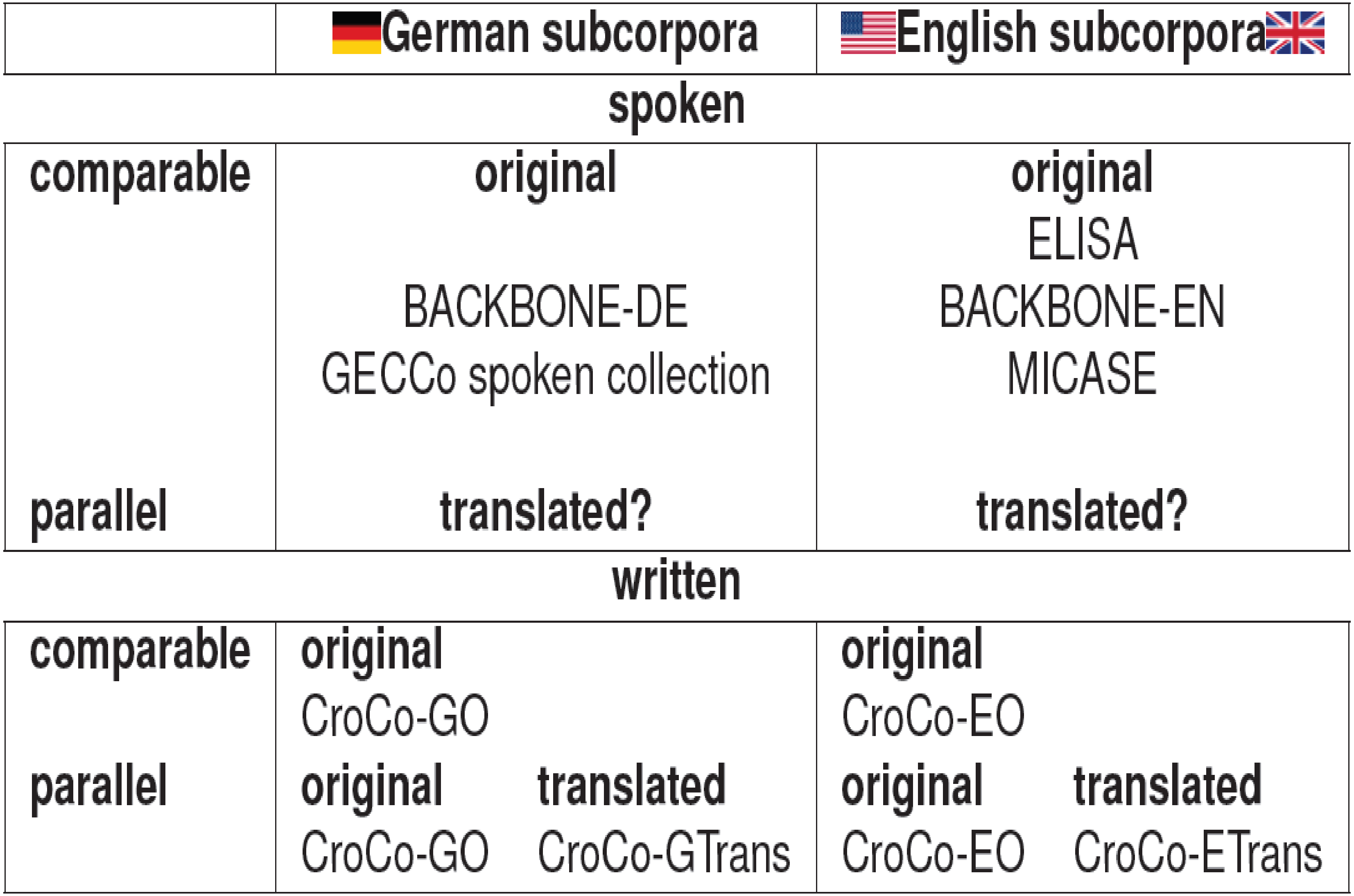
(3) We work for prosperity and opportunity because they're right. It's the right thing to do. [EO_ESSAY_006]

(4)Wir arbeiten für Wohlstand und Chancen, weil das richtig ist. Wir tun damit das Richtige. [GTrans_ESSAY_006]

(Locally) Un-resolvable anaphors and active processing

We just got back from France. **It** was great fun.

(Cf. Eckert and Strube 2000: 57)



	German subcorpora		English subcorpora	
spoken				
comparable	original BACKBONE-DE GECCo spoken collection		original ELISA BACKBONE-EN MICASE	
parallel	translated?		translated?	
written				
comparable	original CroCo-GO		original CroCo-EO	
parallel	original CroCo-GO	translated CroCo-GTrans	original CroCo-EO	translated CroCo-ETrans

2.2.2.: GECCo corpus structure including spoken registers (cf. Amoia et al 2011)

2.2.3. Assumptions (1)

Due to the more differentiated and transparent morphological encoding of the greater part of German grammar, and particularly within the proforms themselves, and within the constraining local (“predicative”) contexts of those forms (cf. Eckert and Strube 2000, Doherty 2002:82ff)), there should be fewer ambiguities in antecedent-proform links in German than in English. In particular, we assume that the *relative proportions of uniquely identifiable to ambiguous to vague to non-identifiable antecedents per pro-form* is higher in English than in German (somewhat related to Eckert and Strube’s 2000: 51 proportion of *NP : sentential : non-identifiable antecedents*). This applies largely to 3rd person, and in particular singular, preforms.

As for the proforms/ anaphors, we assume a version of Ariel’s (2001: 31) **accessibility marking scale**: *Full name+modifier > full name > long definite descriptions > short definite descriptions > last name > first name > distal demonstrative+modifier > proximate demonstrative+modifier > distal demonstrative+NP > proximate demonstrative+NP > distal demonstrative – NP > proximate demonstrative-NP > stressed pronoun + gesture > stressed pronoun > unstressed pronoun > cliticized pronoun > verbal person inflection > zero*

2.2.3. Assumptions (2)

In the German clause, and certainly in the clause-complex, there are greater uncertainties and flexibilities of where a particular quantum of information may go topologically than in English (word order as a direct expression of information structure, so-called “free word-order”)). Hence, there is less clear syntactic markedness and therefore more ambiguity of structural focus, and a large potential of multiple (primary, secondary…) foci (Doherty 2002: 42ff; 82ff in particular). In this respect, we assume that there should be more ambiguity in antecedent-proform links in German, at least in written discourse, where intonation cannot be used for encoding, some of which should take the form of ambiguities of scoping of substitutes and ellipsis. Hence the necessity of focusing particles and also more explicit morphological marking of various kinds in German

2.2.4. Provisional summary of lexical differences between the two languages.

Few differences in the structural organization of content words (König and Gast 2012:258f).

English may tend to some more general words in some lexical fields, and then exhibits fewer selectional restictions than German.

Compounding is more frequent, more productive and internally more complex in German than in English. It is an altogether more clearly grammaticalized category in the former than in the latter (König and Gast 274).

Prefixes, particles and prepositions are more productively used in German verbal derivation (276ff).

Whereas in general, the German pronoun system is more elaborate than the English one, in a few areas (reflexives, indefinite pronouns), English is morphologically more differentiated.

As for focus particles and modal particles, German has more of them and they express more semantic differentiations (König and Gast 2012: 298ff), although there are “local” exceptions to this rule.

Some studies report move towards more lexical specificity in some areas of English-to-German translation (cf. Zinsmeister et al. 2012:77f).

The overall vocabulary of English may well be larger than that of German (cf. among others Leise and Mair 1998: 41ff, 65), and it may be more Romance than Germanic nowadays.

However, in terms of frequency of usage as measured by type-token-statistics or by comparative frequencies of the n-most frequent words, German lexis may in fact be more varied (Leisi and Mair 1998: 46, 65f, as well as our own work in e.g. Steiner 2008; Hansen-Schirra et al 2012).

2.2.5. Quality and sustainability of data:
(1) Inter-annotator agreement

Interannotator-Agreement							
GO				EO			
	A1/A2	A1/Asys	A2/Asys		A1/A2	A1/Asys	A2/Asys
GO_FICTION	0.85	0.78	0.73	EO FICTION	0.92	0.91	0.95
GO POPSCI	0.77	0.61	0.54	EO POPSCI	0.94	0.88	0.85
GO WEB	0.89	0.54	0.56	EO WEB	0.9	0.89	0.96
GO TOU	0.85	0.66	0.63	EO WEB	0.91	0.93	0.96
Mean	0.84	0.65	0.62	Mean	0.92	0.9	0.93
	0.84	0.64			0.92	0.92	

Cf. Amoia 2013 in progress

(2) Precision and recall

Precision and Recall of Automatic Annotation Framework					
GO			EO		
·	Precision	Recall		Precision	Recall
GO FICTION	0.77	0.74	EO FICTION	0.95	0.91
GO POPSCI	0.51	0.65	EO POPSCI	0.89	0.84
GO WEB	0.71	0.44	EO WEB	0.96	0.89
GO TOU	0.63	0.66	EO TOU	0.96	0.93
Mean	0.66	0.62	Mean	0.94	0.89

Cf. Amoia et al 2012, 2013 in progress

2.2.6. Basic types of data and a statistical processing pipeline
(1) Constructions: chains of anaphor – antecedent

(2) Properties of constructions:

2.1. Properties of the “anaphor” (categories from the system networks in our systemic contrasts, e.g. personal vs. demonstrative vs. substitution and subtypes of them; categories from Ariel’s accessibility marking scale)

2.2. Properties of the antecedent (syntactic function, focus status, phrasal status, abstractness, animacy etc.)

(3) **Properties of the link** (=construction), e.g. “endophoric vs. exophoric”, “unique vs. ambiguous vs. vague” and “phrasal vs. supra-phrasal vs. non-identified (where “exophoric” means “no link to antecedent in text”).

(Cf. for some basic ideas Wiechmann 2011; Diversy, Evert, Neumann. to appear)

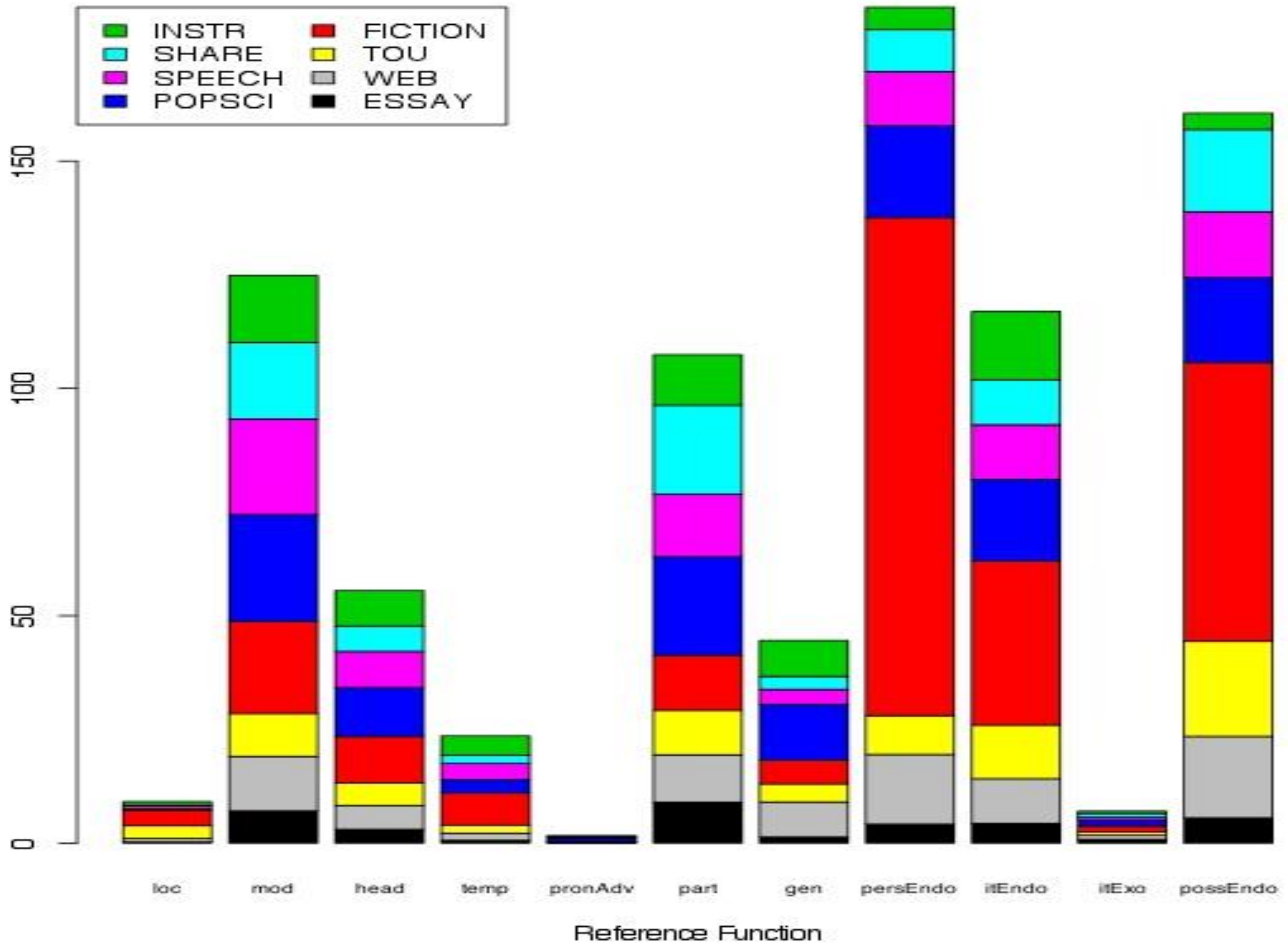
Types of analyses on the data in the GECCo-corpus:

A multivariate contrastive analysis of cohesive reference in English/German (Amoia et al 2012; in progress)

Hypothesis: "*Reference is a distinguishing feature for register variation*".

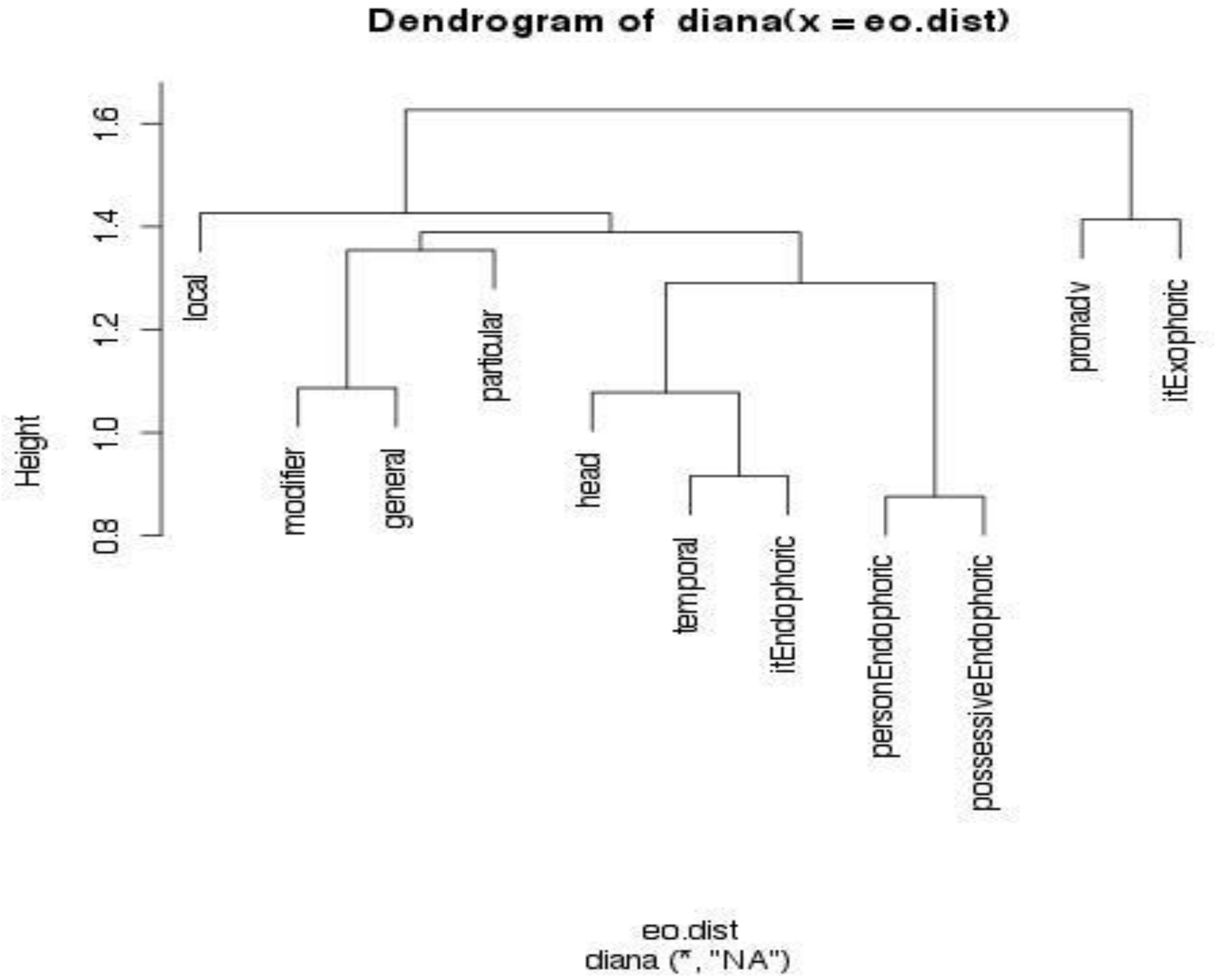
Following (Halliday and Hasan,1976) we distinguish three main types of reference, annotated in the corpus together with their function. Thus the corpus includes annotations of a total of 11 different subtypes of reference (e.g. demonstrative modifier, comparative general, possessive endophoric, etc.), on which a multivariate analysis is performed.
We apply the following **methodologies for quantitative data analysis** to test the hypothesis:

(1) Registers by **devices**: Descriptive Data Analysis (English only),
(2) Registers by **devices**: Clustering,
(3) Registers by **devices**: Supervised Classification
(4) Co-reference **chains** by a set of chain properties: frequencies and interactions



(1) Descriptive Data Analysis: Reference Function by Register (English Originals)

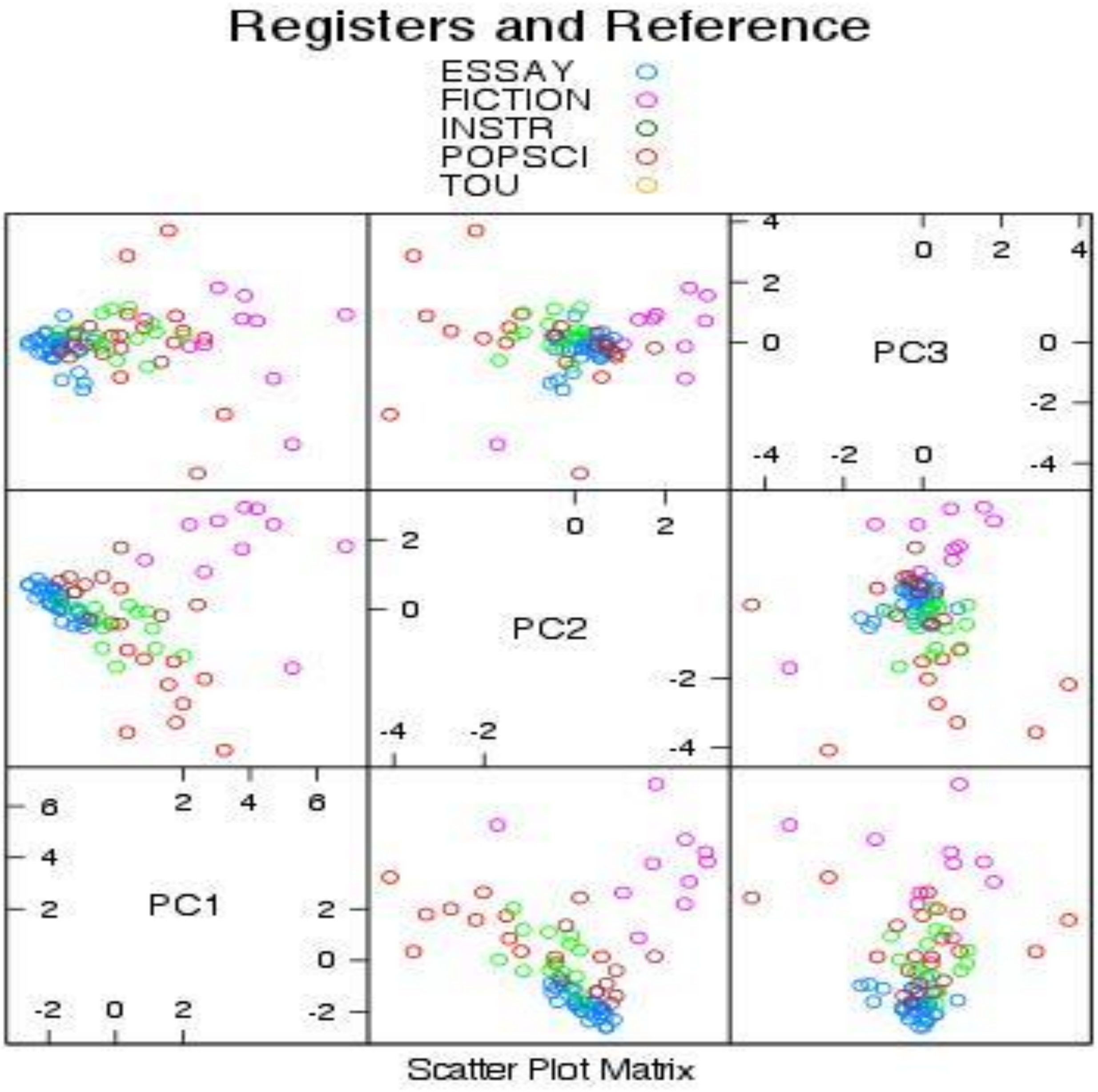
In a second step, we exploited **unsupervised clustering techniques**, in particular principal component analysis, to identify correlations between groups of features. Our clustering analysis below shows that the types proposed by (Halliday and Hasan,1976) have different degrees of strength as descriptors and that further sub-types may be useful. For instance, we found that the personal it-exophoric and demonstrative pronadvs types cluster together in English and form a highly correlated subclass of reference types.



(2) Clustering

The analysis of variance shows that the following 5 features can be regarded as major predictors of register variance:

personEndophoric,
possessiveEndophoric,
general,
local,
itEndophoric



(2) Unsupervised clustering:

In a final step, we applied **supervised classification** techniques such as classification trees and support vector machines (SVM) (cf. (Joachims, 2006), (Karatzoglou et al., 2006)) in order to identify the subset of reference types/features that are distinctive of the different registers and can be used for enhancing automatic register classification.

We found that only a subset (5 out of 8) of the registers included in the corpus can be identified by the set of distinguishing features we used. Interestingly, the registers that the classifiers could not distinguish were those including less cohesive texts such as web pages and political speeches.

The best classification of registers by reference features (classification accuracy 78.87%) is achieved for the 5 registers ESSAY, FICTION, INSTR, POPSCI, TOU (set EOs3) by considering the set of 4 features: personEndophoric, possessiveEndophoric, general, local.

(3) SVM supervised classification – data under revision

Parameters	AccuracyEO	AccuracyEOs1	AccuracyEOs2	AccuracyEOs3
all	50.91%	67.57%	51.28%	66.20%
pred1	49.09%	72.97%	41.03%	76.06%
pred2	48.18%	75.68%	46.15%	78.87%
pred1: personEndophoric, possessiveEndophoric, general, local, itEndophoric				
pred2: personEndophoric, possessiveEndophoric, general, local				
EOs1: ESSAY, FICTION, REGISTER, INSTR, POPSCI, SPEECH				
EOs2: WEB, SPEECH, SHARE				
EOs3:ESSAY, FICTION, INSTR, POPSCI, TOU				

(4) Co-reference chains: Preferred Strategy in written vs. spoken English.

For the analysis we use a subcorpus of English including two text genres POPSCI (11 texts from popular science journals) and INTERVIEW (11 manually transcribed interviews). In order to characterize differences between texts in written and spoken genres we used the following metrics:

T-length, token length of coreferring expressions.

S-distance, the number of sentences separating coreferring expressions in the same chain.

Parallelism, the number of coreferring expressions that exhibit the same or similar syntactic features as the next preceding coreferring expressions in the same chain (e.g. sentence-initial subject).

Cont next slide Amoia 2013

Grammatical role preference of different types of coreferring expressions (e.g. noun phrases occurring as objects).

Typology of coreferring expressions (named entities, noun phrases, personal, possessive and demonstrative pronouns, etc.).

Morphological Features of coreferring expressions (singular vs. plural, 1st, 2nd or 3rd person).

Chain Size, i.e. number of coreferring expressions in one chain.

Frequency of types of co-reference vs. Length of chains

We observed a general tendency of written text to prefer lexical cohesion (e.g. N(named) E(ntity), repetitions) over pronominal reference. (cf. Table 3 below).

In spoken genres on the contrary pronominal reference seems to be the most frequently used strategy.

Spoken language constrains referential elements to span shorter (6.5 sentences) text distance (s-distance between anaphor and antecedent) if compared with written text (8.8 sentences in average).

However, lexical cohesion seems to allow equal length segments in both text genres and generally might involve very long spans of text, circa 12 compared with the 2.5 sentences observed for pronominal reference.

These effects might be probably explained by considering short term memory constraints.

	POPSCI	INTERVIEW
TokenPerSent	46	37
Distribution of Coreference Type		
Lex	70.00%	39.00%
Pron	18.00%	58.00%
This	0.01%	0.02%
s-Distance		
Average	8.8	6.5
Lex	11.3	12.6
Pron	2.5	2.6
This	1.5	2.3

Table 3: Chaining strategies written vs. spoken English (Amoia 2013)
data under revision

Table 4 shows that pronominal reference is preferably used in subject position in both text genres and *rarely* in modifying phrases. Lexical reference strategies, on the contrary, have a wider range of syntactic positions and can occur as subj, obj or as a modifier.

The spoken texts also display a higher number of parallel syntactic constructions in coreference chains if compared to written texts (16% of all constructions in the spoken vs. 6% in the written). In spoken texts, the most frequent parallel syntactic constructions have the structure NP1 VP NP2, where NP1 is a subject and NP2 is an object or a complement. This type of construction is less frequent in the written discourse, where we observe more constructions of the type NP VP1 VP2, where VP2 is expressed by an infinitive, or of the type NP VP PP, where the prepositional phrase contains the coreferring NP.

The comparison of structure diversity across spoken and written texts shows that in spoken texts the diversity is not as rich as in the written ones; this can be partially explained by less variation in the type of coreferential device employed.

	POPSCI	INTERVIEW
Lexical Cohesion		
Subj	41.13%	31.75%
Obj	24.46%	53.90%
Mod	34.41%	14.29%
Pronominal reference		
Subj	86.08%	97.19%
Obj	11.39%	2.81%
Mod	3.00%	0.00%

Table 4: Chains over syntactic functions - **data under revision** -

Contrastive Study of Coreference English vs. German

The following **(Table 5)** shows a contrastive analysis of coreference in German vs. English. Again, we use as a subcorpus the register POPSCI that includes for each language 10 texts from popular science journals. In order to characterize structural differences between the two languages we used the following metrics:

TokenPerSent, the number of token in a sentence.

MentionPerSent, the number of coreferring expressions in the same sentence.

Grammatical role preference of different types of coreferring expressions (e.g. noun phrases occurring as objects).

Typology of coreferring expressions (named entities, noun phrases, personal, possessive and demonstrative pronouns, etc.).

Chain Size, i.e. number of coreferring expressions in one chain.

	EO_POPSCI	GO_POPSCI
TokenPerSent	25.75	25.41
MentionPerSent	2.3	1.57
ChainSize	3.6	15
LexicalCoref	63.00%	72.50%
PronominalCoref	25.80%	20.70%

Table 5: Co-reference chains English vs. German (Amoia in preparation)
data under revision

Lexical Coreference		
	EO_POPSCI	GO_POPSCI
Subj	41.13%	70.42%
Obj	24.46%	13.70%
Mod	34.41%	11.20%

Amoia in preparation - data under revision

Pronominal Coreference		
	EO_POPSCI	GO_POPSCI
Subj	86.08%	88.00%
Obj	11.39%	7.30%
Mod	3.00%	4.70%

Amoia in preparation - data under revision

2.3. The relationship between processing and encoding

(1) Processing studies using key-stroke logging, production histories for translations units (macro-units), eye-tracking, recall- protocolls (cf. Alves et al 2010)

(2) An information-theoretic account based on the hypothesis of “uniform information density” (cf. Piantadosi et al 2012; Levy, Jaeger and others).

Example of tra

Inconspicuous with respect to metaphoricity

verb

TT1: *We firmly believe that business success and social responsibility do not necessarily contradict each other.*

verb

ST: *Wir sind davon überzeugt, dass erfolgreiche Unternehmensführung und soziale Verantwortung sich nicht widersprechen.*

TT2: *We are convinced that successful management and social responsibility are not contradictory.*

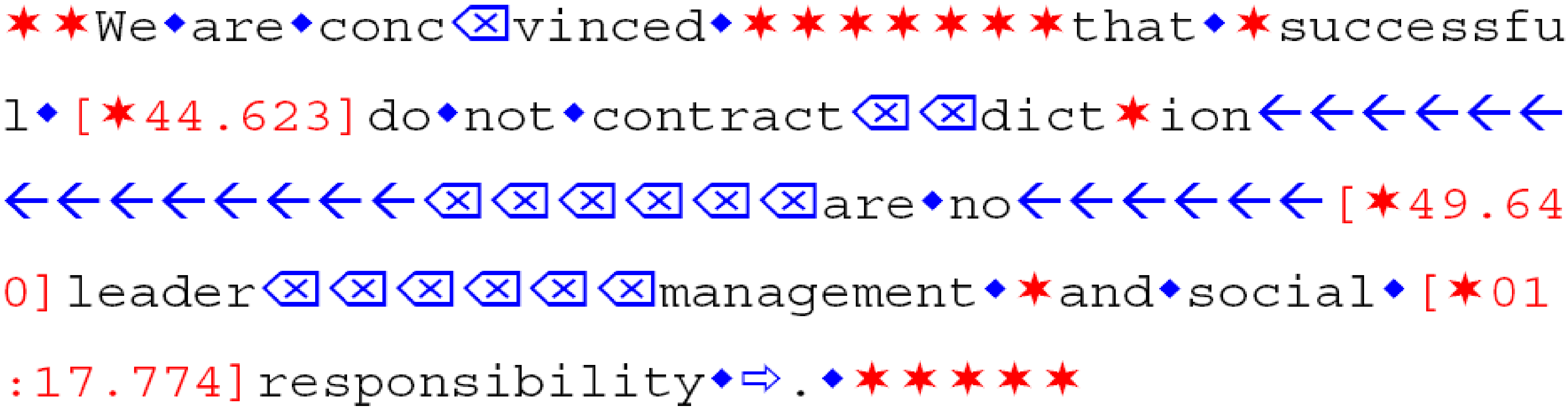
adjective!

Instance of grammatical shift.
→ Explanation?

34

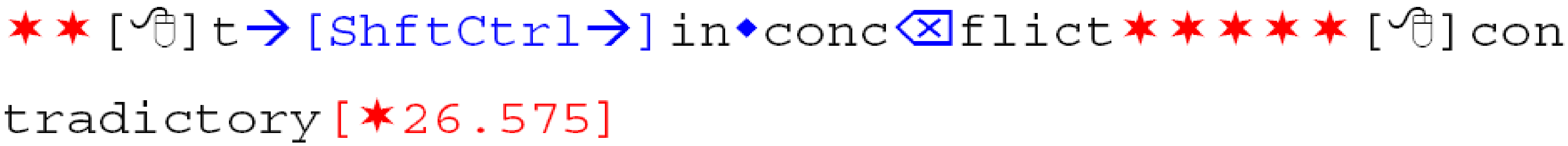
Log file of the translation process

S2: Drafting phase



```
✶✶We◆are◆conc⌫vinced◆✶✶✶✶✶✶✶that◆✶successfu
l◆[✶44.623]do◆not◆contract⌫⌫dict✶ion←←←←←←
←←←←←←←←⌫⌫⌫⌫⌫⌫are◆no←←←←←←[✶49.64
0]leader⌫⌫⌫⌫⌫⌫management◆✶and◆social◆[✶01
:17.774]responsibility◆⇨.◆✶✶✶✶✶
```

S2: Revision phase



```
✶✶[🖱]t→[ShftCtrl→]in◆conc⌫flict✶✶✶✶✶[🖱]con
tradictory[✶26.575]
```

A window on the process

Phase	TT1	TT2
Original	*sich nicht widersprechen*	
Drafting	*are not contradictions in terms*	*do not contradict*
Drafting	*do not necessarily contradict each other*	*do not contradiction*
Drafting		*are no contradiction*
Revision		*are not in conflict*
Revision		*are not contradictory*

Rank shift:
Verb → noun

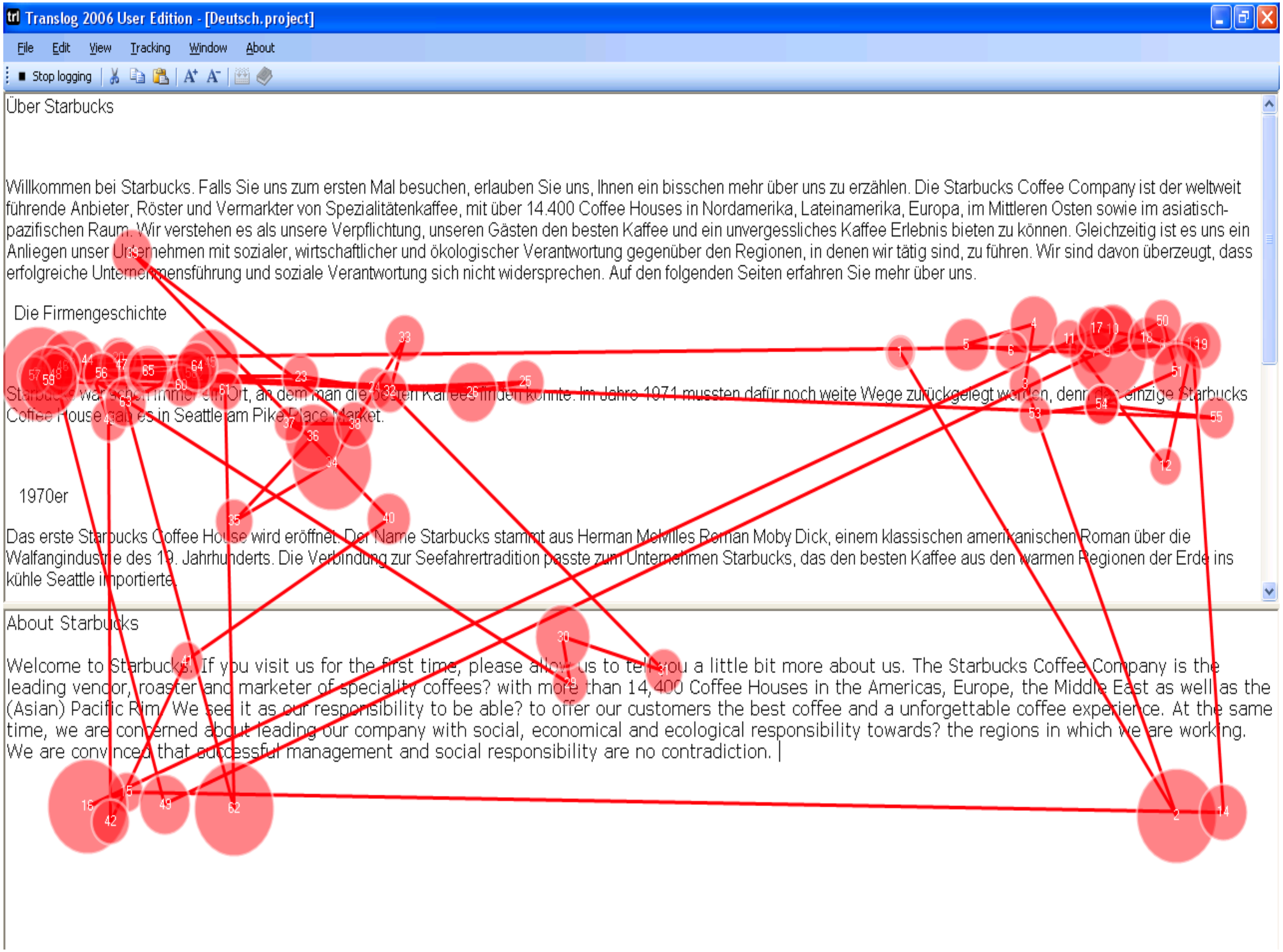
Verb!

Rank shift:
Verb → noun

Back to verb;
effect: no change
in metaphoricity

Rank shift: Noun → adjective;
effect: change in metaphoricity

Insight from eye tracking


Translog 2006 User Edition - [Deutsch.project]
File Edit View Tracking Window About
Stop logging

Über Starbucks

Willkommen bei Starbucks. Falls Sie uns zum ersten Mal besuchen, erlauben Sie uns, Ihnen ein bisschen mehr über uns zu erzählen. Die Starbucks Coffee Company ist der weltweit führende Anbieter, Röster und Vermarkter von Spezialitätenkaffee, mit über 14.400 Coffee Houses in Nordamerika, Lateinamerika, Europa, im Mittleren Osten sowie im asiatisch-pazifischen Raum. Wir verstehen es als unsere Verpflichtung, unseren Gästen den besten Kaffee und ein unvergessliches Kaffee Erlebnis bieten zu können. Gleichzeitig ist es uns ein Anliegen unser Unternehmen mit sozialer, wirtschaftlicher und ökologischer Verantwortung gegenüber den Regionen, in denen wir tätig sind, zu führen. Wir sind davon überzeugt, dass erfolgreiche Unternehmensführung und soziale Verantwortung sich nicht widersprechen. Auf den folgenden Seiten erfahren Sie mehr über uns.

Die Firmengeschichte

Starbucks war schon immer ein Ort, an dem man die besten Kaffees finden konnte. Im Jahre 1971 mussten dafür noch weite Wege zurückgelegt werden, denn das einzige Starbucks Coffee House gab es in Seattle am Pike Place Market.

1970er

Das erste Starbucks Coffee House wird eröffnet. Der Name Starbucks stammt aus Herman Melvilles Roman Moby Dick, einem klassischen amerikanischen Roman über die Walfangindustrie des 19. Jahrhunderts. Die Verbindung zur Seefahrertradition passte zum Unternehmen Starbucks, das den besten Kaffee aus den warmen Regionen der Erde ins kühle Seattle importierte.

About Starbucks

Welcome to Starbucks. If you visit us for the first time, please allow us to tell you a little bit more about us. The Starbucks Coffee Company is the leading vendor, roaster and marketer of speciality coffees? with more than 14,400 Coffee Houses in the Americas, Europe, the Middle East as well as the (Asian) Pacific Rim. We see it as our responsibility to be able? to offer our customers the best coffee and a unforgettable coffee experience. At the same time, we are concerned about leading our company with social, economical and ecological responsibility towards? the regions in which we are working. We are convinced that successful management and social responsibility are no contradiction. |

Recall protocol subject 2

3:34 “… mixture of grammatical and lexical problems … wasn’t sure whether I wanted to use … a nominal or a verbal construction… that’s why I came back to this later”

14:06: “I mentioned that before … I was in conflict (laughter) … I didn‘t like ‚conflict‘ because it seemed too (?) for a corporate text … they wouldn‘t use negative words … that‘s why I changed it back to ‚contradictory‘ I think”

38