

Methodological cross-fertilization: empirical methodologies in (computational) linguistics and translation studies

Erich Steiner
 Universität des Saarlandes, Saarbrücken
 e.steiner@mx.uni-saarland.de

Recent years have seen attempts at improving empirical methodologies in contrastive linguistics and in translation studies through interdisciplinary collaboration with multi-layer corpus architectures in computational linguistics. At the same time, explanatory background for empirical results is increasingly sought in more sophisticated models of language contact in typologically based contrastive linguistics on the one hand, and in language processing in situations of multilinguality, including translation, on the other. Three attempts are discussed to narrow the significant gap between the high level of abstraction of such models, and data provided through shallow analysis and annotation of electronic corpora.

The first of these operationalizes the high level terms “explicitness/ explicitation” in terms of lexicogrammatical data available in a contrastive corpus, treating them as dependent variables and attempting to explain their variation in terms of the independent variables controlled for in the corpus architecture.

The second attempt starts from the same corpus architecture, yet includes annotations about textual cohesion in its operationalizations and develops increasingly fine-grained hypotheses to limit search space and variation between independent and dependent variables so as to get closer to causal explanations rather than explanations in terms of co-variation .

The third attempt intersects corpus data of the type outlined before with data from processing studies, aiming at an integration and mutual explanation of product and process data. Our focus here is on methodological issues involved in integrating data of such different types and granularity in an overall empirical research architecture.

1 Empirical methodologies: some issues to be addressed

Recent years have seen a few, although still limited, attempts at improving empirical methodologies in contrastive linguistics and in translation studies through interdisciplinary collaboration with projects involving multi-layer corpus architectures as developed and refined in computational linguistics. These corpus architectures provide data enriched by a variety of techniques ranging from shallow to deep processing (Vela et al 2007, Čulo et al 2008, Teich et al 2008, Teich and Fankhauser 2010,). They allow the posing of linguistic questions as empirical questions even in areas which until recently were considered the province of hermeneutic debates supported by – at best representative – examples. If such data and their relationship to linguistic theorizing can be clarified, linguistics and translation studies can be made much more empirical than has hitherto been the case (cf. Featherstone and Winkler 2009; Zif 2009, Hawkins 2004 for critical debates in a wider linguistic context).

As a necessary consequence of these developments, empirical methodologies have come under critical scrutiny leading to improved standards of data production, maintenance and analysis. At the same time, explanatory background for empirical results is increasingly sought in more sophisticated models of language contact in typologically based contrastive linguistics (e.g. Thomason 2001, Teich 2003, Doherty

2006, Fabricius-Hansen and Ramm eds. 2008, Siemund and Kintana. eds. 2008, Steiner 2008, Miestamo et al. eds. 2008, Dunn et al 2011) on the one hand, and in language processing in situations of multilinguality, including translation, on the other (Alves et al 2010, Carl et al 2008). The result of these developments is a conceptual and methodological gap between the necessarily high level of abstraction of models on the one hand, and the data provided through shallow (and cheap), or else deeper (and more expensive), analysis and annotation of electronic corpora on the other. It is not immediately obvious where and how stipulated abstract and general properties deriving from models of language variation, contact and change, such as *complexity, explicitness, density, contrast, interference and shining-through etc.* show up in the data, and if so, which of the stipulated independent variables causes which (group of) properties to vary. This gap has to be narrowed through concerted efforts involving methodologies from computational linguistics, including machine translation, (contrastive) linguistics and translation studies, efforts yielding convincing operationalizations of the abstract properties involved. Abstract properties like *complexity, explicitness, density, contrast, interference and shining-through* can thus be linked to patterns in the data available as product data in corpora, or as process data in experimental processing studies.

Beyond this, and quite fundamentally, there is the question of “representativeness” of data: In what sense can we claim that our data, and how much of them, represent the phenomenon we are investigating, rather than some ad-hoc variation caused by any number and kind of independent variables outside the scope of our models. To take just one example, relative explicitness of textual encoding of meaning may be the result of different degrees of context dependentness, of level of subject field expertise (of author and/ or reader), of time-pressure during production, of the dialectics between economy vs. expressiveness, of the degree of training for the production of the register/ genre at hand, of level of education, of formality, of the status of the text produced as an original or a translation etc. etc. If we are interested in the effects of one independent variable, say translation as a mode of text production, we must find ways of isolating it from the other potentially interfering variables. Otherwise, the effects found in our data may be said to derive from something else than the text production mode “translating”.

We shall discuss three test cases of work involving linguists, translation scholars and computational linguists (and marginally psycholinguists): one of them investigates a key notion of translation (*explicitation*) using product-data, the other an under-researched area of language contact (*borrowing and interference phenomena on the level of cohesion*), again using product-data from a corpus, and the third investigates key aspects of language processing during translation, thus focussing on process-data. The gap to be closed exists between the notions of *explicitness/explicitation* and *contact through cohesion* on the one hand, and the level of the available data on the other. If our models of translation, for example, stipulate that translated texts are more explicit than non-translated registerially-parallel (i.e. texts of the same register) original texts in the same language, and if we want to approach this assumption empirically, then we need to operationalize the notions of “explicitness/ explicitation” with respect to the representational categories available in our data. If the categories in our data consist of

- lexical strings,
- annotation layers such as PoS, words, chunks, clauses, sentences,
- statistics on relationships between these,
- alignment phenomena between relevant units in originals and translations such as *crossing lines*, and *empty links*

we need to define, or rather operationalize, the notions of explicitness/ explicitation in terms of these categories, and we need to do so in a theoretically motivated way. Of the categories of data just mentioned the first three should be self-explanatory. *Crossing lines* as alignment phenomena occur when between aligned source-target translation units the source-target link crosses a unit boundary (non-local translations as in the translation of a syntactic subject into an object, or as the translation of a raising-verb into an adverbial). An *empty link* occurs whenever one of the source-target nodes in a translation relationship is empty at a given level of representation.

Seen relative to existing approaches, we are attempting to synthesize individual parameters of language comparison and language contact into more general dependent variables (*explicitness, cohesion*), and we suggest operationalizations in such a way as to enable empirical corpus-based (and ultimately also experimental) investigations. We shall also try to isolate causally related independent variables for the variation observed (section 2). Another attempt at narrowing our search space is the formulation of increasingly fine-grained hypotheses on corpus data as illustrated in section 3. This should allow us to make our observations more precise, and also to systematically reflect textual cohesion, rather than lexis and grammar only. However, this further attempt in itself does not yet solve the problem of uniquely identifying causes and effects. To that end, we shall briefly discuss an attempt at intersecting corpus data with data from processing experiments, in order to find evidence for relationships stipulated by our models of language production, and of translation more specifically (section 4). Finally, an attempt is made to identify achievements as well as persistent methodological weaknesses, and implications are identified for research methodologies.

2 Explicitness of encoding, operationalization in terms of corpus-data and the task of isolating independent variables

The first attempt *CroCo*¹ departed from the assumption that translations as texts are characterized by the property of *explicitness* relative to registerially parallel original texts within the same language. Elaborate tests were conducted on corpora of translations and registerially parallel texts in the target languages English and German. A further assumption was that this explicitness is due to the translation process, taking the form of *explicitation* observable cross-linguistically between source and target text segments, so-called “translation units”. Translation units were then searched for explicitation phenomena causing the observed differences in “explicitness” (cf. Table 2 in section 4). Register and language no doubt both play their parts as independent variables causing variation in explicitness, yet the assumption here was that the translation process plays its own theoretically motivated role in this configuration. The abstract notions of *explicitness/ explicitation* have their own history both in translation studies and linguistics, yet have only rarely been subjected to empirical studies (cf. Englund-Dimitrova 2005 and the literature cited therein).

The *CroCo-corpus* is partitioned into 8 registers each in English and German (cf. Hansen-Schirra et al 2007, Vela et al 2007, Steiner 2008), plus one cross-register reference corpus for English and German each. The sub-corpora were compiled using sampling techniques (Biber et al 1998) and annotated for PoS, morphology, chunks, syntactic functions, clauses and sentences (cf. Culo et al 2008 for an overview of the tools used). The sub-corpora of original and translated texts can be compared along

¹ Cf. <http://fr46.uni-saarland.de/croco/>; funded by DFG 2005 - 2009

all of the annotation layers, including combinations of them, both within and across English and German. A second and important source of data were alignments between originals and their translations on all of the levels annotated (i.e. word, chunk, clause, sentence cf. Čulo et al 2011). Figure 1 shows the corpus structure.

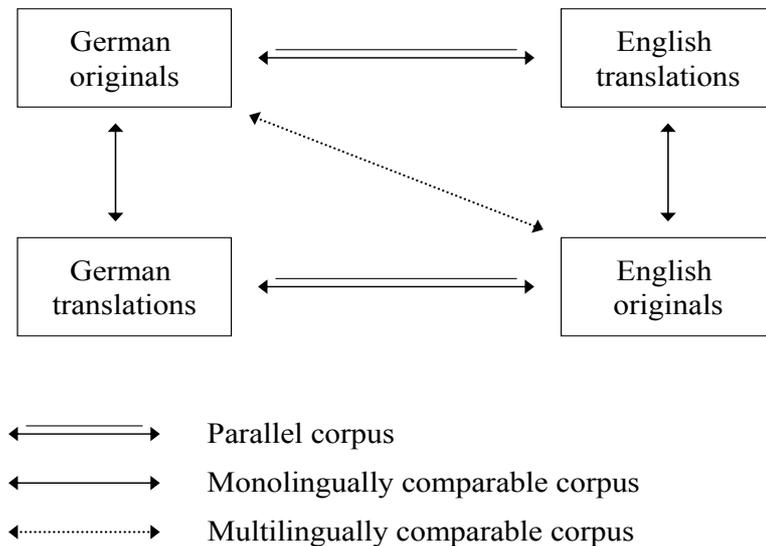


Figure 1: Bidirectional Translation Corpus (from Hansen-Schirra et al forthcoming ch.2)

The notions of *explicitness* and *explicitation* were then given a careful operationalization (cf. Table 1 for “shallow” annotation layers) in terms of the types of information contained in the different configurations of relevant sub-corpora (cf. Figure 1). It was then possible to show *whether* and *to what extent* the phenomenon of “explicitness” obtained for any of the sub-corpora compared to the others.

Table 1 uses as features low level data in the form of lexical density (LD), type-token-relationships (TTR) and part-of-speech tagging (PoS). The contrasts C1-n in the second column refer to contrasts between sub-corpora (reference corpora (ER, GR), corpora of originals (EO, GO), translation corpora (ETrans, GTrans), and register specific corpora within originals and translations as listed in footnote 2. In the third column, we list which indicator(s) in terms of the low-level data we believe to be indicative of which phenomena, and in the fourth column we posit explanations in terms of our three independent variables language, register, and status of a corpus as representing originals or translations.

The independent variables *language system*, *register* and *translation* can be reasonably isolated and related to the observed effects in the data. Remaining questions about representativeness of the sub-corpora can to some extent be approached with future improvements in sampling techniques and corpus size. There is the remaining question of the extent to which our corpora, especially the translation corpora, represent “competent/ standard/ evaluated” translations, rather than data full of opportunistic errors and mistakes. Doherty (cf. e.g. 2002, 11ff; 2006, 1ff and 159ff) strictly defends exclusively “evaluated/ controlled data” as relevant for empirical work. As far as this methodological claim is concerned, we would defend the acceptance of texts as relevant data as long as they have been published as “translations”, our main argument being that judgements about what counts as more or less competent language use are subject to a set of by now well-documented problems in language production generally (cf. e.g. Haider 2009), and in evaluations of translations in particular (House 2001). What our translational corpora do

represent, we would claim, is the language produced in situations culturally accepted as “translating”, which is not at all the same as holding that all these translations are “good” in the sense of being optimized solutions to the general problem called “translation”. Furthermore, if the majority of the translated texts in the corpus can be shown to exhibit the property of “explicitness” relative to original texts, then this property is established as a distinguishing property of this subcorpus – even if in separate evaluations of these translations it can be shown that they are sub-optimal .

Features	Contrast (C1-n)	Phenomenon: Indicator	Explanation
Lexical Density (LD), Type-Token-Ratio (TTR), Parts-of-Speech proportionalities (PoS)	C1 (Reference Corpora ER vs. GR)	- Experiential explicitness: LD (E>G) - Strength of lexical cohesion other than repetition: TTR (G>E) - experiential and referential density: PoS (G>E in nominal orientation)	Language System
PoS proportionalities, reflecting “nominal orientation”	C2.2 (8 Registers within languages E and G)	- Experiential density: nominal orientation	Register, Language
		English: <i>TOU > SHARE > WEB > ESSAY > INSTR > SPEECH > POPSCI > FICTION</i>	
		German: <i>TOU > WEB > SHARE > ESSAY > INSTR > SPEECH > POPSCI > FICTION</i>	
LD, TTR, PoS (Nominal Orientation)		- referential and experiential density: spread of language-internal variation (G>E for TTR and nominal orientation; E>G for LD)	
	C2.1 (EO vs. GO by register, with ER/GR differences factored out)	- experiential and referential density: LD, TTR, PoS	Register
LD, TTR, PoS	C3 (Translations vs. originals within a language and within a register)	- Experiential explicitness: (LD) (ORI>TRANS) - lexical variation: TTR (ORI>TRANS) - referential density: nominality (ORI>TRANS, with exceptions)	Translation Process, De-Metaphorization

Table 1. Summary of “shallow” statistics used as operationalizations for “explicitness” (cf. Hansen-Schirra et al forthcoming ch. 14)²

² Abbreviations in Table 1: Abbreviations are explicitated whenever they occur for the first time in Table 1 if read from top to bottom. Registers are TOU (Tourism), SHARE (Letters to our shareholders), WEB (Websites), ESSAY (Essay), INSTR (Instructions), SPEECH (Speeches), POPSCI (Popular Science), FICTION (Fiction), ORI (Originals), TRANS (Translations)

However, even if it can be argued that a CroCo-type architecture allows systematic studies of co-variation between variables, and even if we make a case for its “translations” to represent relevant data, we have to admit of a significant methodological problem: the third one of our variables, *translation*, if interpreted as *translation process*, is inherently complex and at present still insufficiently-understood (cf. also Becher 2010). Not only does it share all the complexities of monolingual text production, but it is text production under the additional constraints of a source text, plus usually the constraints of a professionally defined situation of production. This methodological problem can be systematically addressed by subjecting the notion of *translation process* to a more detailed analysis and by testing its effects in experimental processing studies involving the cumulation and intersecting of data from key-stroke logging, eye-tracking and post-hoc protocols (cf. Alves et al 2010, see also section 4 below). Arguing on the basis of the results of *CroCo*, we therefore feel justified in claiming that translated texts are characterized by some property, such as explicitness, and that the reason is not either the language, or the register, as these were controlled for separately. However, we are not able to convincingly show *which aspect* of the translation process is related to precisely which sub-aspect of overall explicitness/ explicitation. And finally, it cannot be excluded categorically that two variables, say *translation* (independent) and *explicitness* (dependent) co-vary, but with the causing factor being located outside our model and ultimately causing the co-variation.

As a first evaluation of the *CroCo*-line of research, we would argue that the general corpus-architecture and the data processing employed can be trusted to yield more and also methodologically refined results of the type indicated here, if it is used in replications of our study. But we need improvements in the areas of *modelling* (internally over-complex variables, representativeness of data), *operationalization* of the models in terms of linguistics features, and in *processing techniques* for corpus data (processing pipelines, evaluation and significance of findings) and even more urgently for experimental data to be discussed in section 4 (amount and naturalness of data, experimental design). It is, for example, far from clear which of the product-based frequencies obtained from our corpora are the result of precisely which of the processes observed in eye-tracking or key-stroke logging experiments. There are at present no models known to us which would reliably relate corpus data to data from experiments, at least in translation studies (for a general critique of experimental data and its relationship to linguistic models cf. Schlesewsky 2009). Improvements in modelling can be expected from translation studies and/ or psycholinguistics, better operationalizations should come out of (contrastive) linguistics, and improved processing techniques are under development in computational linguistics. The task at hand now, it seems, is to improve methodologically guided communication between the relevant research communities.

3 Contrasting cohesive patterns in English and German: the role of hypotheses for interpreting corpus data and the challenge of identifying contact phenomena

Our second attempt starts from the same corpus architecture as the one sketched above, yet includes annotations about textual cohesion in its operationalizations and develops increasingly fine-grained hypotheses to limit search space and variation between independent and dependent variables so as to get closer to causal

explanations rather than explanations in terms of co-variation only. GECCo³ sets out from the diagnosis that our current knowledge about English-German contrasts in cohesion is still weak. For contrastive grammar, we have reasonably comprehensive system-based accounts (Hawkins 1986, König and Gast 2007), yet these are not backed-up by empirical validation. Doherty's work (2002, 2006), which we have found very significant in its addressing phenomena of grammar, information structure and some aspects of cohesion, tests what she calls "stylistic" intuitions of competent native speakers and translators (2002,11), based on principles of optimal integration of local textual parts into their relevant discourse context (discourse appropriate translations, Doherty 2006, 1ff). Unfortunately, her test environment is not very controlled and not critically assessed from a methodological point of view (cf. Doherty 2006, ix). Even so, she provides important intuitive and theoretically well-motivated insights into translation. Her overriding goal, however, of testing (previously trained) intuitions, rather than linguistic production and product as such, makes her work methodologically problematic as an empirical investigation.

For cohesion, not even a system-based comparison is available, much less an empirical foundation for such a comparison. The tracing of contact phenomena on the level of cohesion is therefore necessarily still in its infancy (but cf. Hansen-Schirra et al 2007 for an early attempt; Kunz and Steiner forthcoming a,b). Substantial advances in technologies using multi-layer annotated electronic corpora for text-based investigations of phenomena of cohesion hold the promise of placing contrastive accounts on an empirical basis, and beyond this comparison also allow us to trace contact phenomena in suitably configured corpora. A multi-layer representation is used, approaching tree-bank functionality and including aligned data for English and German translations in both directions as a crucial empirical base, with the exception of the spoken subcorpora. Extensive frequency information about cohesive configurations is incorporated into what is essentially an extension and reconfiguration of the *CroCo-corpus* referred to above, tied to varieties or registers of the language concerned, and this time notably including spoken subcorpora (cf. Figure 2).

	 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 translated CroCo-GO CroCo-GTrans	original translated CroCo-EO CroCo-ETrans

Figure 2: GECCo corpus structure including spoken registers (cf. Amoia et al 2011)

³ <http://fr46.uni-saarland.de/gecco/GECCo/Home.html>; currently running and funded by DFG since 2011

The CroCo corpus, partitioned into 4X8 plus two reference corpora, was restructured into 4 subcorpora (GO, EO, GTrans, ETrans) with the registers no longer saved as separate sub-corpora, but as structural attributes of the 4 sub-corpora. For the spoken registers, not contained in the earlier CroCo corpus, the *GECCo-corpus* does not include translations, as these registers are usually not translated. As data for the contrastive work, though, they are sufficient. The new structure allows simpler and faster query in the CQP. Searches in the corpus can still be conducted within a single register or in all registers at the same time. This modified corpus structure implements some improved processing techniques of the type mentioned as desiderata in section 2 above (cf. Amoia et al 2011).

In terms of overall explanations for the data thus obtained, one of the interesting questions is that of whether contrastive properties of cohesion in the two languages point into the same direction as some assumed generalizations in contrastive grammar (directness of mapping from semantics to grammar (G>E), different tolerance of various forms of “ellipsis” (E>G), more explicit encoding in one of the languages in the clause (G>E), possibly the opposite tendency in the verb phrase (E>G), etc.), or whether cohesion serves as a dialectic counterpart, distributing constraints not in the same direction as in grammar, but possibly in the opposite one. In the terms of Hawkins (2004, 44ff), we are ultimately interested in how the two languages cue “processing enrichment” through their different systemic options of cohesion, and ultimately also in whether or not the enrichment, and thus interpretation of discourse units, is differently affected. A further interesting object of investigation are the properties of cohesive (referential and/ or lexical) chains in terms of frequency, length, distance between elements, number and kind of entailments triggered through sense relations in and between lexical chains etc.), which hitherto have hardly been accessible to empirical investigations (but cf. Hansen-Schirra et al. forthcoming for early thoughts along these lines).

Our corpus-linguistic analysis includes the identification of various types of cohesive devices (*reference, substitution, ellipsis, coherence relations, lexical cohesion*; for some important modelling background cf. Halliday and Hasan (1976); Halliday and Matthiessen (2004, 524ff)) and their lexicogrammatical realizations, the linguistic expressions to which they connect (the antecedents), as well as the nature of the semantic ties established and properties of the cohesive chains where appropriate. Including translations in the analysis should provide evidence for analogies between cohesive devices in the two languages, but also show areas where one-to-one equivalents are not preferred, or even non-existent.

The currently existing annotation requires an expansion in terms of additional layers of annotation, which are currently under construction. For instance, particular cohesive devices establishing *reference* or *substitution* can be investigated on the part-of-speech level. Other types such as *conjunctions* can be identified when examining the part-of-speech as well as the chunk level. For the investigation of *ellipsis* combined queries into different layers of annotation can be employed. For the analysis of nominal, verbal or clausal ellipsis the current annotation is too shallow and does not permit a fine-grained differentiation of types of linguistic devices. Thus, more specific cohesive categories have to be developed and annotated.

In order to narrow the gap between the concept of *contact through* cohesion and the level of our data, a structured grid of hypotheses is specified for empirical analysis as a testing ground for

- contrasts in the uses of *similar* systemic resources (e.g. the definite article in German vs. English, or the dependent variable in hypothesis 1 below)
- contrasts in the use of *different* systemic resources for similar cohesive functions/ purposes (e.g. substitution vs. reference through personal

pronouns vs. lexical cohesion for the function of co-reference in German vs. English)

- traces of language contact due to different usages in contact vs. non-contact varieties (categorical and/ or in terms of frequency in comparisons of translated vs. original text of the same register within English or German).

Note that the formulation of hypotheses as such is not a new development in our context (cf. Steiner 1991, 141ff; Teich 2003, 143ff; Hansen 2003, 127ff; Neumann 2008, 89ff), and is, of course, standard practice in many strands of empirical work. What we are using them here for in particular is the motivated narrowing down of search space in our data for the specific purposes of our investigation.

Examples of some hypotheses are:

Hypothesis 1: 3rd person singular neuter pronouns vs. masculine and feminine pronouns (frequency E(nglish)>G(erman) for originals (contrast)), in terms of PoS overall and proportionally within pronouns.

Cf. (1) and (2) for examples from our corpus:

- (1) *Eine verantwortungsbewusste Politik kann diesen Prozess, der zudem von objektiven Faktoren determiniert wird, nicht nur flankieren. **Sie** muss **ihn** vielmehr formen.*
- (2) *A responsible policy can not only accompany this process, which is additionally determined by objective factors, **it** must moreover shape **it**.*

Preliminary tests on hypothesis 1 have been run and are relatively straightforward to carry out with lexical search on our lemmatized sub-corpora. Initial results (cf. Kunz and Steiner forthcoming a) indicate higher overall frequency as predicted, yet sensitive to register, and even unconditioned higher frequency for cohesive vs. non-cohesive usage E>G (cf. Hypothesis 4). Interpretation is less clear, because the observed differences may be due to, at least, the predominance of grammatical vs. natural gender in co-reference for 3rd person singular pronouns in German, the possible preference in German for demonstrative reference over simple personal or possessive reference (cf. Hypothesis 4), the different degrees of availability of lexical cohesion as an alternative to pronominal reference between the languages etc. So, while hypothesis 1 narrows the search space for findings, it does not in itself lead us unambiguously from the observation of co-variation to causal explanation.

Hypothesis 2: GO>ETrans(lations)>EO(riginals) in locally non-ambiguous 3rd person reference within their register.

A German – English contrastive pair of (constructed) texts is given in (3) and (4) below:

- (3) *Mein Freund machte seinen Abschluss, besorgte sich einen Kredit und gründete seine erste Firma. Er/ sie/ es/ der/ die/ das/ dies/ diese(r,s)/ letztere(r/s), der Versuch/ daraus ...wurde ihm zum / entwickelte sich ein Verhängnis.*
- (4) *My friend got his degree, obtained a loan and founded his first business. It/ this/ that/ out of this, the attempt developed (into) a disaster.*

The underlying assumption here is that English translations (versions of (4)) from German (versions of (3)) show less local ambiguity in local antecedent – pro-form relationships than English originals (not exemplified above), inheriting this

(hypothesized) property from their German originals. “Local” here needs to be operationalized into “between adjacent clauses” or some such measure. It can then be tested, if ambiguity is taken to mean “number of possible antecedents for each relevant pro-form”. Our assumptions here are triggered by, once more, the existence of grammatical gender in German, as well as by the higher usage of alternative and less ambiguous forms instead of *es* in German (cf. Hypothesis 4). These findings, if corroborated, should include fewer possible antecedents for German “er/sie/es” than for English “he/she/it”, but certainly fewer possible antecedents for the alternative (demonstrative/ adverbial/ fully lexical) cohesive alternatives. We are referring here to the systemically conditioned availability in German of the demonstrative article, as well as “pronominal adverbs”, both of which have a function of narrowing the range of plausible antecedent phrases for their occurrences if compared with personal or possessive pronouns, providing a motivation for our hypothesis (cf. Kunz and Steiner. forthcoming a: section 4.2.). As far as the cost of this analysis is concerned, we have to trace the possible antecedent – pro-form relationships within a local domain, which as such is possible on the basis of PoS annotations, combined with chunk and clause annotation. Open questions, however, arise out of the fact that co-reference relationships need not be local in the sense just introduced, thus making our measure of “ambiguity”, and in that sense “complexity”, one of local structure of the encoding, rather than an overall textual measure. Nor can we directly infer processing complexity from this local encoding complexity – which has to be taken for granted for any product- rather than process-based work in isolation. Local encoding ambiguities will, in fact, often be tolerated by language producers and processors in the interest of more global efficiency (Hawkins 2004, 47f).

Hypothesis 3: ETrans-T(arget)T(ext)>GO(riginals)-S(ource)T(exts) in explicitated 3rd person reference through use of fully-lexical TT-equivalent of pronominal source

The assumption here is again that German co-reference chains in originals are locally, i.e. between adjacent members of a chain, less ambiguous than in English originals. If this is the case, then one strategy for an English translation would be to use lexically-headed phrases, possibly combined with pre-modifying demonstrative/ deictic material, to achieve a similar effect as their German source text originals. Hypothesis 3 refers to one aspect of Hypothesis 2, the two are thus not strictly independent. In order to obtain the relevant data, we have to retrieve co-referential chains from texts, which at this stage can only be obtained from costly hand-coded small corpora. We also consider chunk-by-chunk alignments between translationally related ST-TT units, which is why we are currently exploring improvements through increased use of tools for lexical chaining. Hypothesis 3 would again successfully limit our search space, however on somewhat costly data, and with a somewhat indirect link to relevant assumptions.

Hypothesis 4: EO>GO in cohesive usage of *it* vs. *es* (because of alternative usage in German of demonstratives of various sorts and pronominal adverbs) between matching registers in original texts, measured both in terms of PoS overall and as proportion of cohesive vs. non-cohesive usage of *it*.

Hypothesis 4 shares some of its background assumptions with hypothesis 1, but in this case we would focus on the use of *it/ es* in cohesive vs. non-cohesive usage. The production of the data is not trivial, though. Our annotation needs to cover grammatically triggered usages of 3rd person singular pronoun *it/es*, because these need to be classified into one relevant sub-class, which would then leave the relevant co-referential and thus cohesive complement class. Again, given the data can be produced at reasonable cost, the hypothesis would successfully limit the search

space, even though the results obtained could be partly due to other interfering factors – though not register or the translation vs. original status, as these are being kept constant.

Hypothesis 5: In terms of the phenomena tested in H1 – H4, we predict that in a comparison of originals and translations (in this case within the same language and register), the translations will diverge from the originals in the direction of their source language.

The background for this explanation is an assumed interference, or rather, shining-through effect (cf. Teich 2003). As some initial findings indicate (cf. Kunz and Steiner forthcoming a: section 4.1.), this is largely, but, dependent on register, not always borne out. Here it will be interesting to trace explanations for why register appears to be an influential variable on some element of the translation process.

Further hypotheses are developed for *comparisons of vagueness/ ambiguity of reference and scope*. Differences can be expected here deriving from usage of different lexicogrammatical realizations of some constant cohesive relationship, or even from different cohesive relationships altogether. An example would be the contrastive use of a generic full lexical phrase vs. a definite phrase vs. a phrase pre-modified through a determiner (possessive vs. deixis vs. demonstrative) vs. a phrase headed by a pro-form (demonstrative vs. pronoun) as tested on aligned ST-TT pairs. The interest would not be in the phenomenon as such, which has been researched under “accessibility rankings” (e.g. Ariel 1990, Hawkins 2004: 45), but in the different kinds of *ambiguity* and/ or *vagueness* associated with each case in interpretation/ enrichment. In general, we would predict that a) translations are less ambiguous and vague than their originals in SL-TL configurations (explicitation through translation), but also b) that they diverge from their original registerially-parallel counterparts in the direction of the respective source language (interference, shining-through).

A final type of hypothesis makes reference to contrastive register-specificity of cohesive configurations, and again their behaviour under contrast vs. contact conditions. For example, German written as opposed to spoken registers may be characterized by dense lexical chains with relatively low lexical repetition, whereas this distinction may be much smaller and involving more repetition for English. For translations from one of these languages into the respective other, we would then predict an interference-like “shining-through” effect (cf. Teich 2003) of source registers onto their target corpora. These configurations will be operationalized as length of lexical or referential chains, density of chains, number of chains per text sample, frequency, length, distance between elements, number and kind of entailments triggered through sense relations in and between lexical chains⁴ etc.. On the basis of WORDNET-type taxonomic classifications, we are investigating different levels of abstraction/ generality in chain progression language internally, but also between aligned lexical translation units. Assuming that it is a frequent translational strategy to resort to a superordinate term as a lexical equivalent in cases of lexical gaps or simply lack of knowledge, one might hypothesize greater generality in translations over originals. On the other hand, if contrastive registers of originals show different degrees of implicitness, possibly realized as higher generality in English of lexically realized concepts, as a register feature, as is sometimes hypothesized in comparisons of English and German texts, this might interfere with translational effects. Add to this the increased reliance of English on “general nouns” as a means of lexical cohesion (Schmid 2000, Mahlberg 2005), and we have grounds

⁴ I am grateful to Marilisa Amoia for emphasizing the importance and accessibility of such relationships to me in recent discussions.

for separately exploring lexical generality as a register feature in originals, and decreasing generality relative to originals in both directions.

Another assumption on which to base hypotheses about lexical cohesion would be that more lay-type registers, rather than expert-type registers, use topological, and often polyphyletic (non-strict inheritance), classification systems rather than typological monophyletic (strict inheritance) ones (cf. Halliday and Martin 1993, 23ff). With the help of WORDNET-based tools for lexical analysis, we can operationalize the concepts of *typology vs. topology* and of *monophyletic vs. polyphyletic* or else *historical vs. genetic*, or *hyponymy vs. meronymy* into lexically-implied sense relationships between elements of lexical chains between registers within and across languages, and between originals and translations. Note that this does not only apply to nouns and their derived adjectives, but also to preferred semantic verb classes: the often observed preference of *relational vs. action* verbs in English over German texts may contribute to generality and thus implicitness of the vocabulary used in lexical chains.

At this early stage of the GECCo-project, we would hypothesize shining-through effects for ST-TT configurations, and for density of chains only a possibly increasing effect of the translation process as such. We need to be aware, though, that the frequency data that can be obtained through work of the type described here is valid and interesting in research on text production in general, whether in monolingual or multilingual contexts, and is furthermore only possible through the joining of efforts from (contrastive) linguistics, translation studies, and computational linguistics.

Where in our research methodology can we trace *contact* phenomena, rather than just *contrasts* in terms of categories and frequencies? In short, where we compare originals of the same register, including the register-neutral reference corpora, across languages, we obtain cohesive contrasts. Where we compare originals and translations within the same language and the same register, any resulting differences would seem to be due to either interference, or else “normalization” in the sense of “hyper-adaptation to target-language norms”. In a weak sense, these are contact phenomena. One possible causal source of these phenomena would then be the translation process, involving some form of “borrowing” (Thomason 2001, 70ff and earlier). Our research architecture is sensitive not only to classical forms of borrowing, but characteristically to shifting frequencies (i.e. over- or underuse relative to the norm established by the same register in the “originals” corpus) below the threshold of structural or lexical borrowing. The translation process in a narrower sense is not the only possible source of contact phenomena in our architecture. The cause of variation could, in fact, be any other component of the contact situation, as long as it impinges on the translation process in a wider sense. In order to make our notion of “translation” more precise, we need to appeal to process studies as shown in the following section.

4 Improving corpus architectures and relating data in corpora to data from processing experiments against relevant models

The third attempt intersects corpus data of the type outlined before with data from processing studies, aiming at an integration and mutual explanation of product and process data. Our focus here is on methodological issues involved in integrating data of such different types and granularity in an overall empirical research architecture. We shall start, though, with a few more general requirements on empirical work of the type discussed here, before concentrating on intersecting different types of data with relevant models.

There is an overall ongoing challenge in research attempts of the type discussed here: The researcher needs to be constantly aware of the cut-off point between very costly “deep” (and to some extent less reliable) annotation, and more “shallow” (and to some extent more reliable) annotation, the latter of which leaves a substantial gap between data and interpretation. Linguistically “deep” annotations, notwithstanding their disadvantages in terms of cost of production and in terms of reliability, have a clearer relationship to highly general models of language processing, whereas the cheaper and often more reliable surface annotations yield data in a very indirect and at worst spurious relationship to more ambitious and general modeling. Our annotation layers in *CroCo* (cf. section 2), for example, involve lexico-grammatical information, some of it shallow and low-cost (part-of-speech-tagging, type-token-ratio, lexical density), some other annotations deeper and involving heavy checking of (semi-)automatic annotations (chunking, clause analysis, and levels of alignments), and some layers even involving annotation by hand requiring monitoring of inter-coder-consistency. Even more challenging in our follow-up project GECCo (cf. section 3), annotations involve those above plus yet more expensive annotations: referential indexing, annotating proform – antecedent configurations, chaining of referential and lexical chains. It is obvious that ways need to be found of producing these with acceptable costs and of sufficient quality, something which cannot be regarded as solved on anything but a small scale. Improved contacts between researchers in translation studies, contrastive linguistics and computational linguistics in particular are essential to make any progress here so as to improve mutual understanding of the issues involved, as well as of the possibilities and limitations of computational technologies available currently.

The question also needs to be raised of how research architectures can be made more standardized than hitherto, allowing independent repetition and (dis-)confirmation of findings. Schlesewsky (2009, 176ff) demands this for experimental data, yet the same is obviously true for corpus data. Relevant research communities need to more systematically share data and replicate each other’s findings in order to arrive at methodological standards comparable to those in the more established empirical research fields. Something like “multicentric studies/ trials” may become possible for some research questions, and possibly most urgently in experimental, rather than corpus-based, studies.

As we have implied in some passages here, and elsewhere (cf. Alves et al 2010), corpora, processing pipelines and evaluated results from corpus-based studies can be used stand-alone as sources of data to check on hypotheses of the types mentioned above. However, they will usually allow the discovery of co-variation of independent and dependent variables only, rather than a necessarily causal relationship. Even if we manage to align source-target units pair-wise within the same register and for only one hypothesis, thus excluding all but one independent variable, we may at best suspect a causal relationship. There is always in principle the possibility that our two variables in independent-dependent pairings co-vary because of some other variable outside our research design, a danger which is more or less plausible, depending on how good our model is. Graded predictions fare somewhat better than categorical predictions, as formulated e.g. in Hawkins (2004, 31ff), yet the basic methodological problem remains, at least as long as the data used are restricted to corpus i.e. product data.

Which brings us to our final point: in order to have a chance of explaining any findings we may have, we need a model, and if at all possible a model predicting the relevant behavior of our variables. The model and its derived hypotheses need to be precise enough to be falsifiable on our data. This is not always the case in (psycho-)linguistic studies generally (cf. Schlesewsky 2009, 170ff), and very hardly at this point in translation studies. And finally, we need to relate corpus data to behavioural data

in the widest sense (eye tracking, key-stroke logging, think-aloud protocols, production time or reaction time studies, EEG studies, FMRI, generally to psycholinguistic and even neurophysiological data) to pave the way towards more principled explanations of the results obtained in corpus studies. This is not because psycholinguistic and neurophysiological data show us the “working of the mind” directly, but rather because they provide additional, and in some cases possibly more direct windows into the mind, even though the latter is not directly observable. Provided, that is, that we have models of translation, language contact etc. which make predictions for the data that we have.

Table 1 above shows data and interpretations from intra-lingual comparisons and inter-lingual comparisons, yet at that stage without any “parallel” corpora, i.e. source-unit into target-units mappings. Assume now that we have such additional data as shown in Table 2 (PoS-shifts in aligned translation units) and Figures 3-5 below⁵:

TYPE OF SHIFT	E-G	G-E
verb-noun	24.31	16.98
verb-adjective	11.69	02.80
verb-adverb	06.95	00.25
adjective-noun	17.43	09.48
adjective-verb	01.84	09.92
adjective-adverb	01.42	11.58
noun-adjective	13.89	21.63
noun-verb	05.74	16.98
noun-adverb	03.40	01.08
adverb-adjective	10.06	01.34
adverb-noun	03.05	01.59
adverb-verb	00.21	06.36

Table 2. Frequencies of PoS-shifts (%) (F. Alves et al 2010, 116)

The data shown in Table 2 are frequencies of PoS-shifts in source-target word alignments (not restricted to the passage shown in Figures 3 to 5), eye-fixations from an eye-tracking study (Figure 3), key-stroke logging data from the same text passage in Figure 4, and process data in Figure 5 showing shifts in intermediate solutions from two subjects translating the passage shown in Figures 3 and 4. In order to interpret these data, we clearly need a type of modelling of the relevant linguistic processes (translation, language production) which makes predictions for these kinds of data. As the situation is currently, we may have models making predictions for the linguistic data, and existing models of translation procedures may even make predictions about shifts as shown in Table 2. Yet our models are still too unspecific – and models about a different domain – to make predictions about eye movements and key-stroke logging directly. The links between cognitive processes in translations and those kinds of data are quite indirect and probably much more prone to interference by other factors, than the purely linguistic data are.

As an illustration of the kind of hypothesis we would suggest here, look at Hypothesis (6) below:

⁵ Project ProBral, funded by DAAD and CAPES 2008-2011

Hypothesis 6: We assume that in producing a given translation unit for a trigger source-text unit, a highly metaphorized (nominalized) passage in comparison to an experientially equal less metaphorized source passage will

1. trigger a higher number of attempted intermediate word-alignments before the final solution is produced,
2. trigger more and/ or longer eye fixations on the problematic unit
3. trigger longer pause units and more attempts plus more revisions in the key stroke units for that passage.

We also predict that the effects are negatively correlated to training of subjects and to length of time given for the task, but positively to direction of translation (into foreign vs. into native language). We furthermore predict a scale of relative strength of these variables training > length of time > direction of translation to be mirrored in relative frequencies of 1. to 3. above.

A window on the process

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

Rank shift: Verb → noun

Rank shift: Verb → noun

Rank shift: Verb → noun

Back to verb; effect: no change in metaphoricity

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

10
10

Figure 5: translation process data showing shifts in intermediate solutions

5 Conclusion and outlook

The significant properties of hypotheses such as our illustrative H(6) above are that it makes predictions for all of our strands of data and that it is based on a *ranking* of independent variables as to strength of effect. We would thus also be looking at *graded effects* in the data, rather than just on yes/ no – effects. But note at the same time that in order to derive hypotheses such as H(6) above, we need models (of the translation process in this case) making predictions in terms of our data. And this is an area where conceptual work needs to be invested: existing models of translation are not fine-grained enough to make this sort of prediction at the moment, so these models need to be developed before studies using combinations of data from corpora and processing data can achieve the effects which they deserve. We are not claiming here that the problems involved are insurmountable, but rather that they are quite general to empirical language studies, and that we should improve communication across relevant research communities to find solutions. Empirical methodologies in contrastive linguistics and translations studies stand a lot to gain from such

developments by being able to become more truly “empirical”. The relevant sub-fields of computational linguistics, on their part, will find much-needed applications for (partly) existing solutions in search of relevant problems, but may even derive intelligent new solutions.

6 References

- Alves, Fabio, Adriana Pagano, Stella Neumann, Erich Steiner & Silvia Hansen-Schirra. 2010. “Translation Units and Grammatical Shifts: Towards an Integration of Product- and Process-based Translation Research“. In: Shreve, Gregory.M. & Erik Angelone (eds.) *Translation and Cognition*. Amsterdam: John Benjamins.109-142
- Amoia, Marilisa, Kerstin Kunz, Ekaterina Lapshinova-Koltunski. 2011. Discontinuous Constituents: a Problematic Case for Parallel Corpora Annotation and Querying. In: *Proceedings of the 2nd Workshop on Annotation and Exploitation of Parallel Corpora (AEPC2 a RANLP 2011 workshop)*. Hissar, Bulgaria. September, 2011.
- Ariel, Mira. 1990. *Accessing Noun-Phrase Antecedents*. London: Routledge
- Becher, Viktor. 2010. “Abandoning the notion of translation-inherent explicitation: against a dogma of translation studies. In: *Across Languages and Cultures* 11 (1), pp. 1–28 (2010)
- Biber, Douglas, Susan Conrad, and Randi Reppen, 1998 *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Carl, Michael, Arnt Lykke Jakobsen, and Kristian T.H. Jensen 2008 Studying human translation behavior with user-activity data. In *Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science, NLPCS 2008, Barcelona, Spain, June 2008*, Bernadette Sharp and Michael Zock (eds.), 114–123. Setúbal, Portugal: INSTICC Press.
- Čulo, Oliver, Silvia Hansen-Schirra, Stella Neumann, and Mihaela Vela 2008 Empirical studies on language contrast using the English-German comparable and parallel CroCo Corpus. In *Proceedings of the LREC 2008 Workshop “Building and Using Comparable Corpora”*, Marrakesh, Morocco, 31 May 2008, 47–51.
- Čulo, Oliver, Silvia Hansen-Schirra, Karin Maksymski, and Stella Neumann. 2011. Empty links and crossing lines: querying multi-layer annotation and alignment in parallel corpora. In: *Translation: Computation, Corpora, Cognition*. Vol. 1. No.1. 2011
- Doherty, Monika. 2002. *Language processing in discourse. A key to felicitous translation*. London: Routledge.
- Doherty, Monika. 2006. *Structural Propensities. Translating Nominal Groups from English into German*. Benjamins, Amsterdam:
- Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson & Russell D. Gray. 2011. “Evolved structure of language shows lineage-specific trends in word-order universals” in: *Nature* 473. 2011: 79-82
- Englund-Dimitrova, Birgitta 2005 *Expertise and Explicitation in the Translation Process*. Amsterdam: Benjamins.
- Fabricius-Hansen, Cathrine and Wiebke Ramm eds. 2008 “Subordination” vs. “Coordination” in sentence and text: a cross-linguistic perspective. Amsterdam, Philadelphia: John Benjamins, Studies in Language Companion Series,

- Featherston, Sam / Winkler, Susanne (ed.). 2009. *The Fruits of Empirical Linguistics. Volume 1 Process. Volume 2 Product*. Berlin: De Gruyter Reihe: Studies in Generative Grammar [SGG] 101
- Haider, Hubert. 2009. "The thin line between facts and fiction". In: Featherstone and Winkler. eds. 2009. Vol.1. 75 – 102
- Halliday, M.A.K. and Ruqaiya Hasan 1976. *Cohesion in English*. London : Edward Arnold
- Halliday, M.A.K. and James .R. Martin. 1993. *Writing Science: Literacy and Discursive Power*. London and Washington D.C.: Falmer Press.
- Halliday, M.A.K. and Christian.M.I.M. Matthiessen. 2004. *An Introduction to Functional Grammar* London: Arnold (earlier versions by Halliday in 1985/1994).
- Hansen, Silvia. 2003. *The Nature of Translated Text. An Interdisciplinary Methodology for the Investigation of the Specific Properties of Translations*. Saarbrücken: DFKI/Universität des Saarlandes.
- Hansen-Schirra, Silvia, Stella Neumann and Erich Steiner 2007. "Cohesion and Explicitation in an English-German Translation Corpus". In: *Languages in Contrast* 7(2): 241-265.
- Hansen-Schirra, Silvia, Stella Neumann and Erich Steiner. forthcoming. *Cross-linguistic Corpora for the Study of Translations. Insights from the language pair English – German*. Series Text, Translation, Computational Processing. Berlin, New York: Mouton de Gruyter
- Hawkins, John A. 1986. *A Comparative Typology of English and German: Unifying the Contrasts*. London: Croom Helm.
- Hawkins, John A. 2004. *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press
- House, Juliane 2001. How do we know when a translation is good? In: Steiner and Yallop. eds. 2001. 127-160
- Kunz, Kerstin and Erich Steiner. forthcoming.a „Towards a comparison of cohesive reference in English and German: System and text , in: Taboada, Maite, Susana Doval Suárez and Elsa Álvarez González forthcoming. *Contrastive Discourse Analysis. Functional and Corpus Perspectives*. London: Equinox
- Kunz, Kerstin and Erich Steiner, forthcoming b. Cohesive substitution in English and German: a contrastive and corpus-based perspective, in: Aijmer, Karin and Bengt Altenberg. eds. forthcoming. *Advances in corpus-based contrastive linguistics. Studies in honour of Stig Johansson*. Amsterdam: John Benjamins
- Mahlberg, Michaela. 2005. *English General Nouns. A corpus theoretical approach*. Amsterdam: John Benjamins
- Miestamo, Matti; Sinnemäki Kaius, and Fred Karlsson. eds 2008. *Language Complexity. Typology, Contact, Change*. Amsterdam/ Philadelphia: John Benjamins
- Neumann, Stella. 2008. *Contrastive Register Variation. A quantitative approach to the comparison of English and German*. Habilitationsschrift Philosophische Fakultät II, Universität des Saarlandes. Saarbrücken
- Schlesewsky, Matthias. 2009. Linguistic data from experimental environments: a multi-experimental and multi-modal perspective. In: *Zeitschrift für Sprachwissenschaft ZFS* 2009. 169-178
- Schmid, Hans-Jörg. 2000. *English Abstract Nouns as Conceptual Shells. From Corpus to Cognition*. Berlin, New York: Mouton de Gruyter

- Siemund, Peter and Noemi Kintana (eds.). 2008. *Language contact and contact languages*. Amsterdam: Benjamins (Hamburg Studies in Multilingualism Vol. 7).
- Steiner, Erich. 1991. *A Functional Perspective on Language, Action, and Interpretation. An Initial Approach with a View to Computational Modeling*. Berlin, New York: de Gruyter
- Steiner, Erich. 2008. "Empirical studies of translations as a mode of language contact - "explicitness" of lexicogrammatical encoding as a relevant dimension." in: Siemund, Peter and Kintana, Noemi. eds. 2008. *Language contact and contact languages*. Amsterdam: John Benjamins (Hamburg Studies in Multilingualism Vol. 7). pp. 317-346
- Steiner, Erich and Colin Yallop eds. 2001. *Exploring Translation and Multilingual Text Production*. Berlin, New York: Mouton De Gruyter
- Teich, Elke 2003. *Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts*. Berlin, New York: de Gruyter.
- Teich, Elke, Richard Eckart and Monica Holtz (eds). 2008 Workshop Linguistic Processing Pipelines. Technische Universität Darmstadt, 10 July 2008, Darmstadt, Germany. <http://www.linglit.tu-darmstadt.de/index.php?id=abstracts>
- Teich, Elke and Peter Fankhauser. 2010. Exploring a corpus of scientific texts using data mining, in Stefan. T. Gries, M. Davies and S. Wulff (eds), *Corpus-linguistic applications. Current studies, new directions*, Rodopi, Amsterdam, pp. 233–248.
- Thomason, Sarah G. 2001. *Language Contact. An Introduction*. Washington D.C.: Georgetown University Press.
- Vela, Mihaela, Silvia Hansen-Schirra, and Stella Neumann 2007 Querying multi-layer annotation and alignment in translation corpora. In *Proceedings of the Corpus Linguistics Conference CL 2007*, Birmingham, UK, 27-30 July 2007, Matthew Davies, Paul Rayson, Susan Hunston, and Pernilla Danielsson (eds.). http://ucrel.lancs.ac.uk/publications/CL2007/paper/97_Paper.pdf.
- ZfS *Zeitschrift für Sprachwissenschaft* 2009: Linguistic data: acquisition – evaluation – theoretical implications