THE *HOWS* AND *WHYS* OF CODING CATEGORIES IN A LEARNER CORPUS (OR "HOW AND WHY AN ERROR TAGGED LEARNER CORPUS IS NOT *IPSO FACTO* ONE BIG COMPARATIVE FALLACY")

KARI TENFJORD*, JON ERIK HAGEN**, HILDE JOHANSEN***
University of Bergen

Abstract

A new electronic learner corpus of annotated language test essays of Norwegian as a Second Language developed at The University of Bergen is presented in outline, and in this connection the question whether the practice of error recording and error coding in itself is theoretically misguided by virtue of the so-called "comparative fallacy" argument (Bley-Vroman 1983) is discussed. Error tagging is defended on the basis of a broader discussion of relations between theory and data in SLA generally, and the overall conclusion is that error recording and error coding is not methodologically misguided since error analysis is not a theory of SLA but rather a method, a method that can, in principle, service any theory. We think that the so-called "comparative fallacy" charge often levelled against classical error analysis or any preoccupation with errors at all stems from failure to distinguish sufficiently between these two notions within SLA.

Keywords: learner corpus, comparative fallacy, error, error tagging, SLA methodology

INTRODUCTION

This article will present an electronic corpus of written test essays produced by learners of Norwegian as a second language known as the "ASK corpus"[1] ("ASK" being acronymic for the three constituent morphemes of Norwegian "andrespråskorpus"), its design and scientific potential. The main aim of constructing this corpus was to enhance the facilities for empirical studies on the acquisition of Norwegian as a second language and perhaps SLA studies more generally. The corpus will provide a database not only of the essay texts themselves, but also annotations of various kinds such as parts of speech, morphological categories and error tags. These resources in combination with an efficient user interface system, make it possible to test hypotheses generated by previous studies in Norwegian as a second language and it may also be a rich source for explorative

---

[1] For a presentation of ASK visit : http://spraktek.aksis.uib.no/projects/ask

*Associate professor Kari Tenfjord, **associate professor Jon Erik Hagen, ***graduate student Hilde Johansen. We are all affiliated to the Dept. of Scandinavian Languages and Literature, University of Bergen.

studies to generate new hypotheses of lexical, grammatical and textual features of written SLA, as well as hypotheses on individual and external factors influencing the language acquisition process in more general terms.

We hope, in addition, that by presenting the ASK corpus, clarifying matters of what at the outset might seem fairly basic and trivial, such as exactly what we are coding, what coding categories we have found practicable and how we go about our actual tagging, we may also with due modesty shed some light upon broader issues of a more general kind, issues concerning basic relations between data and theory of SLA in any language on a general epistemological level.

Our aim is therefore twofold: not only to present the basic architecture of our corpus itself, its general outline, purpose and process of development but also to argue that this sort of corpus of error tagged interlanguage texts is a perfectly legitimate aid to SLA research. This second point is occasioned by our having been challenged by quite natural and reasonable questions like "How do you handle IL phenomena A, B and C, etc. in your corpus?", questions motivated by the more or less general underlying assumption that any sort of tagging of interlanguage (IL) texts will, unless it is based on a thorough theoretically motivated pre-analysis of the individual essay texts themselves, be open to the charge of the so-called "comparative fallacy" in SLA (i.e. Bley-Vroman 1983).

THE COMPARATIVE FALLACY IN SLA

The presentation of our corpus, and the observations and reflections following from it will hopefully demonstrate that the notion of "error" in SLA research is not an intrinsically misguided concept. Error recording and error coding does not necessarily presuppose a theoretical model of second language acquisition implying that seeking insight into second language competence is merely a matter of performing systematic comparison between learner language and target language in terms of the structures and entities of the target language only.

As Bley-Vroman (*op.cit.*) quite rightly points out this sort of thinking fails to appreciate the autonomy and integrity of the interlanguage, and the patterns and forces governing its existence, use and development and the complexity of factors influencing it at various stages. This would be a theoretical approach to SLA which more or less assumes acquisition of items of the target language, and that language acquisition is viewed as a sort of metaphorical journey from L1 to L2 with various stages of "interlanguage" as intermediate stations between these terminal points.

One noteworthy point of progress of SLA research is that although the field of SLA-research is a heterogeneous one, and the aims, methods and theoretical perspectives vary considerably, there seems to be virtual universal agreement that "the comparative fallacy" in fact *is* a fallacy. The basic premise that the autonomy of the interlanguage must be recognised and cannot be considered merely some sort of derivative of either L1 or L2 is more or less taken for granted these days

Basically the argument is this: If the theoretical perspective upon second language acquisition process is limited to systematic analytic comparison between TL and IL, and IL is analysed in terms of the structure and categories of TL, then the researcher will fail to appreciate the autonomy and integrity of the interlanguage, and will be unable to describe its inherent communicative functions, its structural independence of both L1 and L2 and its autonomous development patterns.

The basic problem, as formulated by Year (2003) in a reference to Bley-Vroman 1983, is the misguided assumption that recording and classifying errors in L2 performance in conjunction with pre-defined analytical procedures more or less "reads off" the interlanguage competence of the language learner. Year's article provides several instances of this pitfall in a variety of SLA studies.

Lakshmanan & Selinker (2001) elaborate this point further, asserting essentially -- our interpretation -- that this link between deviance from L2 norm and its description and the interlanguage competence underlying that performance is by no means a straightforward one.

This issue became quite a challenge to us, since systematic error tagging involves precisely the sort of comparison Bley-Vroman (1983), Selinker & Lakshmanan

(2001) and Year (2003) warn against. Nevertheless we think that our corpus is a valid methodological aid to SLA for reasons that will be apparent below.

EXPLICIT DESIGN CRITERIA FOR THE ASK DATABASE

The raw material of the ASK corpus consists of a finite number of test essays produced by candidates sitting for two tests in Norwegian as a second language, "Språkprøven" ("The language test") and "Test i norsk - høyere nivå"("Test in Norwegian - higher level"). In terms of *The Common European Framework of Reference for Languages*, these tests measure language proficiency at two distinct levels, B1 (Threshold level) and B2 (Vantage level) respectively. These tests are produced, managed and administrated by "Norsk språktest", a public body affiliated to, but organizationally independent of, The University of Bergen. Our data have been sampled from the their archives, which represents a unique source for building a corpus.

An obvious advantage is that these test essays are reasonably homogeneous as far as the circumstances of their production are concerned. Text genre and context of production are constant for a large number of candidates varying in age, social background, L1, education, etc. In order to enhance this homogeniety further, the sample essays have been collected from the same test situation for every test taker. They have all had the same time frame sitting for the test, they have taken the test under the same conditions, the tests have been scored by assessors with the same kind of training and the personal data of the candidates have been recorded in conjunction with the same test occasion.[2]

Nevertheless, a limit to the actual size of the corpus had to be set for a variety of practical reasons, and reasonably informed decisions how to select our data from this large source had to be made.

From the very beginning, our aim was to construct a corpus that would represent

---

[2] The Norwegian Language Test institution is doing serious research on rater reliability of The language test and Test in Norwegian - higher level (Carlsen 2003, Moe and Jones 2003)

an entirely novel opportunity to perform quantitative research into Norwegian as a second language based on much larger samples than had hitherto been possible, and we hope it will be useful for the entire community of Norwegian second language research scholars irrespective of their subfield and theoretical orientation. We hasten to add, however, that there will of course be limits to what kind of research questions our corpus will be a relevant resource for. One such obvious limitation, for instance, is the fact that the ASK database contains written material only, and cannot therefore, at least not directly, be of use for those interested in oral data.

Our design critera are motivated by the following considerations. We wanted a corpus of written texts produced by a group of learners

a) from a broad range of mother tongues. ("the L1 criterion")

b) with a certain minimal *documented* degree of communicative proficiency ("the PASS criterion")

c) for whom we would have access to a variety of personal data potentially relevant to any aspect of their second language profile, be it language structure, use, development or proficiency or any other conceivable point of interest potentially reflected in the texts produced ("the INFO-criterion").

The learning context is of course not quite the same for every candidate, and the individuals differ in various ways. But since we have coded information of important personal variables we can control for differences both in learning context and in learners' background. We are confident that the corpus passes the criterion of corpus utility proposed by Granger (2002:9): "The usefulness of a learner corpus is directly proportional to the care that has been exerted in controlling and encoding the variables".

The L1 criterion

The basic criterion for selecting texts for the corpus is the mother tongue of the learner. This is a theoretically motivated criterion since the effect of this variable has been widely discussed in the field of SLA: whether the mother tongue (L1) has any

effect on second language acquisition, and if so, in what way does it affect language learning? Today there appears to be widespread agreement among SLA researchers that L1 affects the learning process in some way or other, but we are facing methodological problems in testing hypotheses concerning the role of the mother tongue. Isolating the L1 factor from other factors which influence language learning may not be possible. A promising methodological approach, in our view, is to carry out statistical analyses of the language produced by learners with different mother tongues while keeping other factors constant for the learners. This methodology will, at least to a certain extent, be possible when performing SLA research based on the ASK corpus.

A second concern for selection was the need for variation in language typology. This criterion did however compete with another issue: the number of texts from learners of various L1s available. We estimated that in order to obtain enough data for statistical analysis of L1 influence on SLA within the resources available, our corpus should contain 100 essays on each level, composed by learners of the same L1. This consideration has, as a consequence, reduced somewhat the degree of typological variation; it has not been possible to find as many as one hundred essay texts from the two different language tests in all the source languages that we would have chosen if we could choose freely. The languages chosen are the following: German, Dutch, English, Spanish, Russian, Polish, Bosnian-Croatian-Serbian, Albanian, Vietnamese and Somali.

In addition to the different L1 groups of learners, our corpus also contains textual and personal data from native Norwegians; 100 informants will take each of the two tests. These native speakers must to some degree reflect the individual variation among the immigrants. We have therefore chosen informants from groups where we expected a variation in age, sex and educational background (for example choirs and church groups).

The PASS criterion

The written performance of the candidates has been assessed to be at or above certain reasonably well-defined levels of language proficiency.

For both of the two tests, the central criterion of assessment is *communicative functionality*. A successful candidate must be able to communicate the contents of his intentions to a degree that is adequate in terms of the definitions specified at each relevant level. This requirement is extremely important. It is not only a practical issue, but is in fact essential to the methodological validity of our entire enterprise, which will be discussed further below, both in conjunction with our discussion of error tagging and our broader discussion of "the comparative fallacy".

The INFO criterion

In addition to the test answers themselves and the assessment of them, each candidate has supplied salient points of personal information potentially relevant to research in SLA and bilingualism, such as mother tongue, age, duration of residence in Norway, the extent of formal language instruction received an so on. Since our corpus contains this sort of information for all the candidates, external variables of this kind can be correlated with specific linguistic ones.

GLOBAL INTERPRETATION AND RECONSTRUCTION OF TEXTS PRIOR TO TAGGING

Each individual essay will, in what follows, be termed a "text" for convenience of exposition. The texts are read and interpreted as accurately as possible in terms of semantic content. One basic assumption underlying the project was this: In order to use these kinds of texts as data for research into systematic SLA, we must arguably presuppose that they express a reasonably clear, identifiable and coherent content, and moreover that, this content is intelligible and processable directly "on line" by any native speaker with native Norwegian as their only linguistic resource.

Systematic IL analysis of the text is not required for interpretation, and is even explicitly forbidden in our error tagging instructions. The test candidates are, of course, writing in accordance with their interlanguage competence, but their interlanguage *performance* has been certified as communicatively adequate by virtue of the fact that the candidates have passed the proficiency test, with their largely global and synthetic (rather than specific and analytic) test criteria. As noted above, this criterion of having passed the test is a fundamental condition for each essay to be included in our set of data.

Of course, for any text, whether it is produced by a native speaker (or in this case, more accurately, a native writer) or not, the following will always apply: A reader will, as already noted, interpret a text not merely upon the basis of its literal content, but also upon the expectations motivated by its contexts in its broad sense. Both the immediate linguistic context, the broader situational context and the ultimate cultural context in which the essays where written are relevant for accurate interpretation. We therefore feel that considerations of this sort are legitimate to exploit in our interpretation, reconstruction and tagging decisions, unlike systematic interlanguage analysis prior to tagging.

Since the candidates have in fact passed the tests, we are justified in assuming with reasonable confidence that their communicative intentions, though formulated in their individual interlanguage, the structure of which the tagger is in principle totally ignorant, *are intelligible to any native speaker of Norwegian.* So, although the interlanguage may be structurally or lexically different from the target language, it does function as a serviceable tool of practical communication for the candidate.

Identifying errors in language use is a meaningless exercise if the content of the utterance is not assumed fairly clear. This fundamental insight is the very basis of Corder's algorithm for classical error analysis (Corder 1973:256 - 294). And our test candidates have all been certified by virtue of their test score to be able to make themselves understood, i.e. understood by readers (or listeners) with no other linguistic resources than native Norwegian language competence.

Lakshmanan & Selinker's (2001) point that transcripts may not provide an adequate rendition of the author's communicative intention, may thus at least to some degree have been met.

The basic caveat for our annotation of the texts is therefore that we are assuming methodologically that our native reconstruction of each sentence based on global interpretation will in most cases reflect the semantic intention of the writer. There may of course be problems when more than one such reconstruction seems possible. Ideally, of course, all alternative reconstructions should be recorded, and the text annotated with different sets of tagging in accordance with each alternative. This, however, would not be a viable solution from a practical and financial point of view. Instead, we have opted for the following two basic principles, the principle of pragmatic probability (abbreviated as "PP" below) and the principle of minimal modification (abbreviated as "MM" below).

> a) The principle of pragmatic probability (PP):
>
> Select the interpretation *most probable* from a pragmatic point of view, taking into account the interpretation of the entire text and the broader context in which the text was produced.
>
> b) The principle of minimal modification (MM):
>
> Select the reconstruction deviating least dramatically from the original; in other words: opt for the reconstruction involving minimal alteration of the original text.

These two principles are not always compatible, and may present certain dilemmas, but this is in our view inevitable irrespective of what strategy we might select.

Let us consider the following example, where the PP takes precedence over MM: In Norwegian, the sequence of attributive adjectives immediately followed by their head noun, i.e. the syntactic structure Adj+N, must not be confused with compound lexical nouns with adjectives as first compound constituent. In the former case, spacing between adjective and noun is obligatory, whereas a compound noun (unlike in English) does not allow space between its subconstituents at all.

This difference is also reflected in the prosody of native oral language, and is of great semantic significance, the compound having a more specific and less

analytically transparent meaning than the syntactic two-word structure. Thus the expressions "rødvin" and "rød vin" (both meaning "red wine") are semantically distinct; the former is a special sort of wine (also known as "claret" in English) which is most often, but by no means always, red in colour. Other sorts of wine, such as rosé wine, cannot be called "rødvin" no matter how red they might be in colour. Similarly "rød vin" on the other hand can refer to any sort of wine, clarets or otherwise that happens to be of the colour red, irrespective of other properties. Thus the two sentences:

"Jeg liker rød vin" =I like (any sort of) wine red in colour

"Jeg liker rødvin" = I like claret (of unspecified colour)

are both well-formed and in principle both may occur in Norwegian, each with its own distinct meaning. From a purely technical and pragmatically blind point of view, no error tagging is in principle warranted and thus, in accordance with MM, no error should be recorded. Yet we know from general practical experience that splitting compounds is a common error produced by learners, and a pragmatic assessment of the situation would suggest that having one's likes and dislikes of wines depend on their colour rather than their taste is a less probable interpretation in most contexts than the "rødvin" interpretation. From this sort of reasoning, it seems natural and legitimate to tag "red wine" as an "oversplit" error, in spite of the fact that the more far fetched interpretation is possible in principle.

The tag applied by our system in cases like these (abbreviated technically "SPL") thus means: "Provided that our reconstruction of "rød vin" as "rødvin" motivated by general pragmatic considerations is justified, the learner has split a compound noun thus violating the target code in this respect".

Note that in this case PP takes precedence over MM, since our semantic interpretation is not based upon mere processing of the utterance itself but also upon reflections on pragmatic probability.

When the errors are of a purely formal kind, and the potential alternatives differ in formal structure only, MM is applicable. Consider the following example:

*Denne retter bruker vi flest i vårt daglige livet

="these dishes use we most in our daily life"

(i.e. these dishes most of us use in <u>our daily life</u>)

The erroneous structure "*vårt daglige livet" has several possible (synonymous) structural reconstructions of which the following may be of interest here:

TR1: Disse rettene bruker de fleste av oss i vårt dagligliv

TR2: Disse rettene bruker de fleste av oss i vårt daglige liv

TR3: Disse rettene bruker de fleste av oss i dagliglivet

These alternatives are all equally plausible. Although they are structurally rather different, they are quite synonymous. So in this case, since there are no semantic issues involved, the principle of minimal alteration will be applicable, and TR2 would be the appropriate choice by that criterion, since the only modification required would be the selection of the inflectional form "liv" instead of "livet", but it might be argued that while TR2 requires less modification, TR1 is slightly more idiomatic. But any scholar finding this sentence amongst his data, would (assuming she is a native speaker of Norwegian) immediately recognize the structural (but not semantic) ambiguity here; so although there is a certain risk of arbitrariness inherent in our systems, this is not catastrophic, since the transparency will in nearly all instances be obvious to any researcher analysing the data provided by searching our database. In this case, the sentence could not be counted as a reliable instance of "SPL" error.

On the basis of pragmatic assessment, and by application of native competence of Norwegian as the only resource, a reconstructed version of each text is produced, which constitutes the basis of error classification.

ANNOTATIONS OF TEXTS INVOLVING IDENTIFICATION OF ERRORS DO NOT IN THEMSELVS CONSTITUTE A THEORY OF INTERLANGUAGE STRUCTURE

When a text has been properly interpreted and mentally reconstructed by the tagger in accordance with the principles outlined above, and on the basis of this interpretation a concomitant targetlike reconstruction of it has been produced, it is

possible to identify the errors contained in it and categorize them. Recall that a clear interpretation of the sentence is a logical precondition for reconstructing it. Whenever there are passages that are too obscure to interpret properly, no errors are identified, and the whole passage is marked with the tag "X" (meaning "Unidentifiable error").

The specific points where the reconstructed text (produced in accordance with the principles discussed above) deviates from the IL text, are carefully noted and defined meticulously and explicitly in terms of grammatical and lexical categories relevant to the description of the *target* language. Note that what are tagged are *not* grammatical categories as such, but rather the type of deviation from the target each error represents assuming *for purely analytical purposes only*, that the portions of the text that do not appear to violate target language grammar or idiom are ascribed an analysis dictated by the grammatical and lexical system of the target language. We have for instance no qualms about identifying parts of speech in the texts in accordance with the rules of the target language. This analysis is thus searchable quite similar to monolingual corpora. It is worth emphasizing at this juncture that the corpus is not just a corpus of errors but a resource of texts annotated by grammatical and lexical categories as well as errors.

Our assumption that using the target language categories for describing the texts allows us to define sub-types of errors, consisting of mere deviations, where words of certain types are either deleted, added or shifted in various ways. For the sake of illustration we provide a complete list of our error categories here:

List of error categories

Lexical:	
W	(wrong word)
ORT	(orthographic error)
PART	(overcompounding)
SPL	(oversplitting)
DER	(deviant derivational affix used)
CAP	(deviant letter case (upper/lower))
FL	(Non-Norwegian word)
Morphological:	
F	(deviant selection of morphosyntactic category)
INFL	(deviant paradigm selection, but interpreted to be in accordance with the morphosyntactic category in Norwegian)
Syntactical:	
M	(word or phrase missing)
R	(word or phrase redundant)
O	(word or phrase order)
The deviation category O has the following subcategories:	
INV	(non-application of subject/verb inversion)
OINV	(application of subject/verb inversion in inappropriate contexts)
MCA	(incorrect position for main clause adverbial)
SCA	(incorrect position for subsidiary clause adverbial)
Punctuation:	
PUNC	(wrong selection of punctuation mark)
PUNCM	(punctuation mark missing)
PUNCR	(punctuation mark redundant)
Unidentified error:	
X	(impossible to interpret the writer's intention with the passage)
The coding categories F, CAP and PUNC have the following subcategories:	
AGR	("agreement errors" i.e. errors following logically from, and triggered by, previous errors, the agreement itself being in accordance with the target language norm)

We are, of course, certainly not suggesting by any means that this more or less mechanical procedure will yield an intrinsically valid description of the interlanguage system itself, i.e. the candidate's mental grammar, so to speak. Though it is possible to envisage any specimen of learner language as containing two sorts of entities, "correct language" and "errors", this would truly be a "comparative fallacy", in the sense of Bley-Vroman (1983). What we are doing instead is to ascribe to the text, or if you will, *impose* upon the texts, a structural description that is reasonably well-defined, yet theoretically neutral.

The complete tagging is not in itself a structural description of the learner interlanguage, and for that very reason, the TL description imposed upon the text by

the analyst provides an objective method for pointing out phenomena that any adequate theory of SLA must be able to account for on a theoretical level. Our procedure merely provides a tool for perception, identification and classification of errors, i.e. language usage deviating in definable ways from the way native speakers use the target language.

Error tagging has been established as a standard procedure in learner language corpora, a consequence of the fact that this kind of corpora needs its own techniques:

> "... computer learner corpora quite naturally call for their own techniques of analysis [...] such as error tagging, which are specially designed to cater for the anomalous nature of learner language" Granger (2002:18)

The techniques being developed must however be in accordance with the relatively modern insight that learner languages have their own autonomous linguistic structure not derivable from L1 or L2. We therefore emphasize that the terms 'error' and 'error coding' are not intrinsic features of the learner language but mere technical terms defined empirically and pre-theoretically with no particular theory of learner language or language learning processes being presupposed. We therefore want to stress as clearly as possible that 'errors' are not inherent entities of the interlanguage as such, but rather analytical concepts imposed upon the texts by us, in order to procure systematic data that any valid theory of SLA should be able to account for.

Error identification is thus a practical rather than a theoretical task, and error analysis is a *method*, not a theory of SLA. This categorical distinction has often been confused in the literature, and allegations that any sort of method involving error identification and classificatio and TL comparison in general will run into theoretical problems of the kind pointed out by Bley-Vroman, makes sense only if this distinction is not made. Theory and method must not be confused. As early as 1973 Corder (1973) made this point abundantly clear in the following passage, where the focus is upon the IL sentence *there is bus stop*, where the obligatory indefinite article has been "omitted" by the learner:

> "The omission of the article (in this case) is only the surface evidence for an erroneous or idiosyncratic linguistic system. A full description of the error

involves 'explaining' it in terms of the linguistic processes or rules which are being followed by the speaker [...]

Of course, superficial description is a necessary condition for linguistic explanation but it is not a sufficient one [...] (Corder 1973:277)

What has been tagged in our corpus is precisely this sort of superficial annotations (the word "superficial" not being derogatory). We are making no claim at all that we are tagging inherent and valid categories of the interlanguage as such, or indeed that we are making any descriptions or theoretical statements about the inherent grammatical structure or content of the learner language whatsoever, for the simple reason that the categories for which we are tagging the texts, are not those found in the IL grammar in the same way as those found in a monolingual corpus of tagged texts produced by native speakers.

OBSERVED FACTS AND THEIR POSSIBLE EXPLANATIONS

We have often been asked questions like "How does your corpus handle transfer data?", "How does your corpus handle text genre differences", "How does your corpus handle miscategorizations, overgeneralizations" and so on. Our simple answer to all of these questions are: "It doesn't". For in our view, "transfer", "genre differences" and "overgeneralizations" are not *explanandum* terms (which are relevant to our tagging) but rather *explanans* terms which it is the task of any proposed theory of SLA to provide. This vital point may be illustrated by a simple example; learners of Norwegian with English as their mother tongue quite frequently produce phrases like

   *"det hus"   (=the house)

   *"den dag"   (=the day)

   *"den kirke"   (=the church)

i. e. phrases with a pre-posed definite article which is in fact found in Norwegian, but it never occurs unstressed immediately before its head noun. It can only occur before an intervening attributive adjectival phrase or quantifier of some sort. The

correct equivalent to the English structure of definite article immediately followed by its head noun is instead the noun in its special inflectional form. The TL equivalent of the English expressions "the house", "the day" and "the church" is thus respectively

"huset"

"dagen"

"kirken"

Any first term undergraduate student of Norwegian as a second language would probably identify a form like "*det hus" as an obvious case of "transfer" since it so transparently mirrors the corresponding English form "the house" or as an "overgeneralization", since Norwegian in fact does have pre-posed definite articles like English, but their use is more constrained than their English counterpart. The brightest amongst them might even observe that the phenomenon could be an instance of the so-called "multiple-effect" principle in the sense of Lakshmanan & Selinker, Gass & Selinker and others. These forms are *not*, however, tagged as either "transfer forms", "overgeneralization forms" or "multiple effects", since these are terms belonging to potential theories for *explaining* our data at some theoretical level or other, whereas our job as database developers is to *present* data and make explicit exactly *how*, but not *why*, they deviate from the target language grammar *in terms* of the grammatical description of the latter.

In the ASK corpus the word "det" in "det hus" would be coded as "R" (=redundancy), i.e. "obligatorily deleted" and the word "hus" with "F"(= deviant morphosyntactic form). However, both "det" and "hus" are words occurring in Norwegian and will be automatically tagged for part of speech in accordance with the target language lexicon. Thus any scholar wanting to test a hypothesis on how learners with English as their L1 express definite reference in their learner language, will be able to search our corpus for the relevant deviations or similarities supporting or disconfirming whatever his hypothesis on this issue might be. Our aim is to be neutral as far as theory is concerned. Notice that whether we believe the error is caused by "transfer" or "overgeneralization" or neither or the combined effect of both of these, our tagging of the error would be exactly the same.

Our job is to present *explanandum* data as clearly and succinctly as possible, so that the phenomenon may easily be identified independently of purely theoretical *explanans* notions purporting to explain them.

Another example example might clarify this issue further.

IL sentence: *I dag jeg dro til byen.

English translation: Today I went to the town.

TL reconstruction: I dag dro jeg til byen.

We observe that the IL-sentence differs from the TL-sentence by the non-application of the rule of subject-verb inversion, and the IL-sentence is tagged for having this identifiable feature, where the feature is identifiable because we allow ourselves to apply TL categories to the learner language. We think that any good theory of SLA should be able to make some sort of prediction as to why deviating sentences like *I dag jeg dro til byen* occurs. Yet again, *our* job is not to provide such a theory, but instead to find a useful format to record data relevant for such a theory, whatever its content.

Subject verb inversion in Scandinavian languages as L2 has in fact been investigated by several scholars of quite diverse and partly incompatible theoretical inclinations (i.e. Hammarberg & Viberg (1977), Håkansson & Nettelbladt (1992), Lund (1997), Glahn et. al. (2001) to name but a few. For all of these studies our tagged corpus would have been of potential utility precisely because it is theoretically neutral in its conception. This principle of theory neutrality has been a vital concern to us, and constitutes the basis of our corpus construction.

THE EPISTEMOLOGICAL BASIS OF OUR APPROACH

The design of the user interface of the corpus is of extreme importance, and careful and explicit documentation of this feature is a question not merely of user-friendly practicability, but also one of valid methodology at the scientific level. This follows

from the trivial fact that the way the individual scholar will choose to make use of the corpus, will depend entirely on what is the focus of her specific research project is, what sort of hypothesis she wants to test or what sort of data of a more explorative kind she is seeking, and this may or may not be served by our corpus. In principle, our corpus is intended to aid researchers of any kind and of any theoretical orientation. But of course, whether or not our corpus will be found useful or relevant for any specific research project is a theoretical issue that must be decided upon by the scholars involved in each case.

The central point here is that our corpus is a tool for data collection and data analysis only, and a vital concern for us in our preparation of it was to distinguish clearly between these functions on the one hand, and theoretically motivated research on the other. These two activities must not be confused, and is basic to the argument we are trying to present below.

A central epistemological premise for us has been the following: Before any phenomenon can be properly explained, it must be properly identifiable and describable independent of any particular theory and must be without purely theoretically motivated entities. This is, in our view, a basic precondition for empirical testing of any theory within any field of learning whatsoever (cf. Popper 2002). And the purpose of our tagging system is to do exactly that, to provide systematic descriptions and classifications of phenomena, errors included, that we think any theory of Norwegian as a second language, or indeed SLA in general, should be able to explain to us. Our set of tagging categories is thus not in itself a theory, but rather a *model* (in the Popperian sense) of representing and presenting salient facts. Our system is thus a catalogue of linguistic categories, including identified errors, and provides a systematic description of them. But "errors" are just relations between observable and describable language performance at surface level on the one hand and the target language on the other, they are *not* inherent features or properties of the interlanguage as such. And as a model, in contradistinction to theories, it can neither be confirmed nor refuted, but only found applicable or inapplicable depending upon the empirical import of the particular theory to be tested and the research design of the scholar testing it.

In our project and in accordance with, and inspired by a Popperian philosophy of science no attempt is made to analyze the IL grammar as such. Our only goal has been to provide an objective pre-analytic description of surface structures in terms of target language categories and to identify errors and categorize them in terms of this model of description without any theoretical bias at all as to why these errors are found or what the explanation behind them might be.

We feel -- and this is an epistemological point on our part -- not a mere technical one -- that, unless some sort of empirically defined deviation from the target language norm is properly identified, and a pre-theoretical description is provided for them to make classification possible, no empirical validation of any theory of SLA, or for that matter any other field of learning whatsoever, is feasible. It must be possible, at least in principle, to observe, describe and refer to what any good theory of SLA should explain without making any pre-theoretical assumptions about the content of such an explanatory theory. This is our way of formulating Popper's so-called "demarcation principle" distinguishing scientific propositions from non-scientific ones. It is therefore of paramount importance that the model is an applicable device, a "metalanguage" of description if you will, independent of theoretical terms that makes sense only within specific theoretical frameworks; otherwise it will be quite easy to overlook critical features and counter-examples to one's hypothesis. Popper has much to say about "theory imbued observations" as a serious source of error. For that reason the *explanandum* must be identifiably independent of any particular *explanans* in order for it to qualify as a scientific claim.

THE TL GRAMMAR AS A MODEL FOR EMPIRICAL OBSERVATION

We contend that the target language grammar is the obvious candidate for such a role as descriptive model. What we further assume is that it must be possible, at least in broad outline, to interpret the semantic content of the texts with no other resource than our target language native competence. We contend that this is a necessary

precondition for any valid IL analysis, presuming some elementary, natural and presumably non-controversial philosophical principles of a pragmatic nature.

The results of our tagging are thus a systematic description of surface structures, including identification and classification of structures deviating from the native norm. It makes no empirical claims as to what causes these differences. Our object is solely to prepare data, or rather potential sources of information, to make the testing of interlanguage theories easier and more expedient.

THE UNFEASIBLE ALTERNATIVE APPROACH

A theoretically valid alternative approach could be to perform an exhaustive interlanguage analysis of all the texts, thus providing a more or less complete interlanguage grammar for each text generating the text sentences, and then tag the text in accordance with the units and structural entities presupposed and motivated by such a (more or less exotic) theory. We are not denying that such a project would be theoretically valid, but a project of that kind would in the first place demand more resources than will ever be available in any foreseeable future. A much more interesting objection to such a strategy, however, is that the practical utility of it would in fact be extremely limited. Theories of SLA are so diverse, so controversial and so transient that a database construed on the basis of any one controversial theory would be dated before its construction was completed and it would be less versatile than ours for a variety of reasons.

We feel that the great advantage of our approach is that our descriptive model -- the target language grammar -- is well known and familiar not only in broad outline but in subtle detail by all taggers, all of them being master students of Norwegian as a second language. And above all, describing the errors of all the texts in accordance with the target language grammar provides a unitary framework common to all the texts, which is of obvious methodological advantage in itself, both as far as validity and reliability of tagging is concerned.

WHY THE ASK CORPUS IS NOT ONE BIG COMPARATIVE FALLACY

Since we feel that our enterprise essentially is a pretheoretical one, we will plead "not guilty" to any potential charge of "comparative fallacy" levelled against the ASK project. We feel that Bley-Vroman's methodological argument against any *theory* of SLA seeking to mold interlanguage theory into the procrustean bed of the target language system is fundamentally valid one, and this is perhaps a point upon which there is broad agreement nowadays.

For that reason, however, to exclude *all* kinds of systematical comparison between TL performance and IL performance as a *method*, would truly not only be throwing out the baby with the bath water: it would be to commit a serious category mistake at the conceptual level. Since the ASK corpus with its annotations is not a theory of SLA, it cannot be refuted by Bley-Vroman's argument.

We feel -- though that is another story -- that this failure to distinguish properly between the two notions of *method* on one hand and *theory* on the other has been something of a vexed issue in the short history of modern SLA research. Contrastive analysis, for instance, is strictly speaking just that, a methodological tool of *analysis* potentially useful for obtaining relevant insights. Moreover, we feel that as part of a methodological research design, it is still potentially relevant. What *is* dated, of course is the status of Contrastive Analysis considered as a *theory* of SLA. In the same way, even the so-called "morpheme studies", a controversial SLA method of research reported in a series of works by Dulay & Burt, for example Dulay & Burt 1974, may potentially still be a relevant method of research provided that the data procured by these methods are properly interpreted by a viable theory, i.e. that an identifiable and tenable link between the data provided by such a methodology is possible. The quite valid critique against the theoretical frameworks within which CA, morpheme studies and error analysis respectively were practiced in their heyday, must not be confused with the potential utility of these techniques as methodological aids in themselves. Whether or not these techniques are relevant for validating a theory depends, in our view, not on inherent properties of these

techniques themselves, but rather upon the content of the theory/hypothesis for which it is put to service. In this respect, methodological procedures are not valid or invalid, but rather relevant or irrelevant, depending on the content and empirical import of the hypotheses tested. When Schachter (1986) demonstrated convincingly that error analysis had a serious *limitation* (a better word than "error" which she herself uses) as she suggested in testing out hypotheses addressing avoidance strategies in second language usage, this was in our view a very important insight. She did not in any way invalidate error analysis as a methodological tool or as part of any conceivable research design in general, only that the practical utility of this method was reduced for her particular purpose.

We think that the traditional failure to distinguish between theory and method in the field of SLA has caused quite a lot of unnecessary controversy. The ASK corpus is essentially a methodological tool, and is neither a theory of SLA in itself, nor does it presuppose any particular theoretical notions *a priori*. For that reason making sensible use of the ASK corpus is not -- in our view -- incompatible with the very important insight of the comparative fallacy principle unless one is of the rather dogmatic persuasion that any reference to the target language at all, be it theoretically or methodologically motivated, is thus incompatible -- a position we find untenable.

In fact our methodology of our corpus construction has been designed to avoid this pitfall from the very outset. And we hope that our database will be useful for future research into Norwegian as a Second language

CONCLUDING REMARKS

Some of the established terms in SLA are for example "acquisition", "fossilization", "avoidance" and "overgeneralisation". These are all terms emanating from a way of thinking that reflects and presupposes TL comparison, they are all related to target language-inherent categories; it is not only "error" that may fall victim to this way of thinking. Upon reflection we find that perhaps the entire metaphorical structure of

even modern SLA terminology more or less reflects TL comparison as the very term

interlanguage itself suggest a journey towards a target…

References:
BLEY-VROMAN, R. (1983). The Comparative Fallacy in Interlanguage Studies:
The Case of Systematicity. In *Language Learning. A Journal of Applied Linguistics* Volume 33,1, pp. 1 - 17.

CARLSEN, C. 2003). *Guarding the Guardians. Rating scale and rater training effects on reliability of scores on oral test of Norwegian as a second language* (unpublished Dr.art. thesis, University of Bergen).

CORDER, S. P. (1973). *Introducing Applied Linguistics.* Harmondsworth:Penguin Education.

COUNCIL OF EUROPE (2002). *The Common European Framework of Reference for Languages.* Cambridge:Cambridge University Press.

DULAY, H., BURT, M. (1974). Natural sequences in Second Language Acquisition, *Language Learning 24*, pp 37-53.

GASS, S., SELINKER L. (1992). *Second Language Acquisition:An Introductory Course.* Hillsdale NJ:Erlbaum.

GLAHN, E., HVENEKILDE, A., HAMMARBERG, B. (2001). Processability in Scandinavian Second Language Acquisition, *Second Language Acquisition*, 23, pp. 389-419.

GRANGER, S., HUNG J., PETCH-TYSON, S. (eds.) (2002). *Computer learner corpora, second language acquisition and foreign language teaching.* Amsterdam and Philadelphia: John Benjamins.

HAMMARBERG, B., VIBERG, Å. (1979). The place holder constraint, language typology, and the teaching of Swedish to immigrants, *Studia Linguistica* 31, pp. 106 - 133.

HÅKANSSON, G. (2001). Tense Morphology and verb-second in Swedish L1 children, L2 children and childern with SLI, *Bilingualism: Language and Cognition*, 4,1, pp. 85-99.

LAKSHMANAN, U. & SELINKER, L. (2001). Analysing interlanguage: how do we know what learners know?, *Second Language Research*, 17, pp. 393-420

LUND, K. (1997). *Lærer alle dansk på samme måde? En lengdeundersøgelse af voksnes tilegnelse af dansk som andetsprog.* Copenhagen:Special-pædagogisk forlag.

MOE, E, JONES, N (2003), Using multi-faceted Rasch analysis to validate a test of writing, Acta Didactica:1, pp. 110-127

POPPER, K. R. (2002). *The Logic of Scientific Discovery* (Revised edition:Original 1959). London: Routledge.

SCHACHTER, J. (1974). An error in error analysis, *Language Learning*, 27, pp. 205 - 214.

TENFJORD, K. (2004). ASK – A Computer Learner Corpus In P. J. Henrichsen (ed.) *CALL for the Nordic Languages. Tools and Methods for Computer Assisted Language Learning* Copenhagen Studies in Language 30, Samfundslitteratur: pp.147–158.

YEAR, J. 2003(?): Instances of the Comparative Fallacy, *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, 4,1