Contrastive Linguistics: Theories and Methods

Volker Gast

1 Introduction: The subject matter of contrastive linguistics

Narrowly defined, contrastive linguistics can be regarded as a branch of comparative linguistics that is concerned with pairs of languages which are ‘socio-culturally linked’. Two languages can be said to be socio-culturally linked when (i) they are used by a considerable number of bi- or multilingual speakers, and/or (ii) a substantial amount of ‘linguistic output’ (text, discourse) is translated from one language into the other. According to this definition, contrastive linguistics deals with pairs of languages such as Spanish and Basque, but not with Latin and (the Australian language) Dyirbal, as there is no socio-cultural link between these languages.

More broadly defined, the term ‘contrastive linguistics’ is also sometimes used for comparative studies of (small) groups (rather than just pairs) of languages, and does not require a socio-cultural link between the languages investigated. On this view, contrastive linguistics is a special case of linguistic typology and is distinguished from other types of typological approaches by a small sample size and a high degree of granularity. Accordingly, any pair or group of languages (even Latin and Dyirbal) can be subject to a contrastive analysis.

This article is based on the (intermediate) view that contrastive linguistics invariably requires a socio-cultural link between the languages investigated, but that it is not restricted to pairwise language comparison. Even though it is not a branch of applied linguistics, contrastive linguistics thus aims to arrive at results that carry the potential of being used for practical purposes, e.g. in foreign language teaching and translation. As it provides the descriptive basis for such applications, its research programme can also be summarized as ‘comparison with a purpose’ (E. König). The ‘objective of applicability’ is also reflected in the fact that contrastive studies focus on the differences, rather than the similarities, between the languages compared.

As a first approximation, the method of contrastive linguistics can be represented as in Diagram 1 (for ease of representation, the following discussion will concentrate on pairwise comparison). ‘A(L_n)’ stands for the analysis of a language L_n and ‘A_c(L_1 ↔ L_2)’ for the contrastive analysis of two languages L_1 and L_2.

![Diagram 1: Contrastive linguistics between language-particular analysis and application](image-url)
The schema given in Diagram 1 will be refined below. In particular, the role of ‘bilingual linguistic output’ will be integrated into the picture. This output not only provides the empirical basis for contrastive studies but also functions as a conceptual link between the linguistic systems investigated, as it can be used to establish comparability between categories from different languages.

After providing a brief historical overview of contrastive linguistics in Section 2, Section 3 will address some fundamental methodological issues, in particular the question of cross-linguistic comparability. In Sections 4 and 5, two major types of comparison will be illustrated, i.e. comparison of purely formal categories (consonants) and comparison of linguistic categories that carry meaning or function (tense). Section 6 will deal with generalizations across functional domains (Wh-question formation and relativization). Section 7 will conclude with some remarks on the empirical basis of contrastive linguistics (specialized corpora).

2 Historical remarks

The programme of contrastive linguistics was instigated by Charles Carpenter Fries from the University of Michigan in the 1940s. Fries (1945: 9) contended that “[t]he most effective materials [in foreign language teaching] are those that are based upon a scientific description of the language to be learned, carefully compared with a parallel description of the native language of the learner”. Some years later, this project was put into practice by Fries’ colleague Robert Lado (1957), who provided a comparative description of English and Spanish. The assumption that foreign language teaching can be improved by comparing the learner’s native language with the language to be learned came to be known as the “Contrastive Hypothesis”. Its main assumptions can be summarized as follows (cf. König & Gast 2008: 1):

- First language acquisition and foreign language learning differ fundamentally, especially in those cases where the foreign language is learnt later than a mother tongue and on the basis of the full mastery of that mother tongue.
- Every language has its own specific structure. Similarities between the two languages will cause no difficulties (‘positive transfer’), but differences will, due to ‘negative transfer’ (or ‘interference’). The student’s learning task can therefore roughly be defined as the sum of the differences between the two languages.
- A systematic comparison between mother tongue and foreign language to be learnt will reveal both similarities and contrasts.
- On the basis of such a comparison it will be possible to predict or even rank learning difficulties and to develop strategies (teaching materials, teaching techniques, etc.) for making foreign language teaching more efficient.

The contrastive hypothesis in the form summarized above soon turned out to be too optimistic. It was too undifferentiated in many respects and neglected important parameters of second language acquisition (e.g. natural vs. mediated, sequential vs. simultaneous, second vs. third language, etc.). Moreover, the contrastive programme lacked a solid foundation in learning psychology and was never even put on a reasonable empirical basis, insofar as the intention of producing comprehensive comparisons of language pairs was never convincingly realized. The enterprise of improving foreign language teaching on the basis of pairwise language comparison was therefore abandoned before long, even though a certain plausibility of at least some of the basic assumptions made by early contrastive linguistics can hardly be denied (cf. Kortmann 1998).
New impetus was given to pairwise language comparison in a number of publications from the 1970s and 1980s that did not primarily pursue didactic purposes (e.g. König 1971, Rohdenburg 1974, Plank 1984). These authors regarded contrastive linguistics as a “limiting case of typological comparison” (König 1996: 51) which was characterized by a small sample size and a high degree of granularity. This typologically oriented approach culminated in John Hawkins’ (1986) monograph *A comparative typology of English and German – Unifying the contrasts*. It was one of Hawkins’ primary objectives to reveal correlations between properties of specific grammatical subsystems (esp. syntax and morphology), with the ultimate goal of ‘unifying the contrasts’. Moreover, Hawkins aimed at providing explanations for the correlations he observed and related his contrastive analyses to theories of language processing (e.g. Hawkins 1992). Even though Hawkins’ hypotheses and generalizations met with criticism (e.g. Kortmann & Meyer 1992, Rohdenburg 1992), they provided important insights and helped establish contrastive linguistics as a type of language comparison that was interesting and worthwhile in itself, without pursuing any specific objectives related to second language acquisition or other linguistic applications.

The 1980s and 1990s witnessed a certain diversification in the field of contrastive linguistics insofar as new topics came into the focus of attention (e.g. pragmatics and discourse studies, cf. House & Blum-Kulka 1986, Oleksy 1989), and new empirical methods were introduced, esp. corpus-based ones (cf. Section 7). The availability of specialized corpora (parallel corpora and learner corpora) also led to a renewal of the link between contrastive linguistics and linguistic applications, e.g. insofar as insights gained from (quantitative) contrastive analyses turned out to be useful for translation studies (see e.g. Johansson 1998a).

Most contemporary studies published under the label of ‘contrastive linguistics’ follow the spirit of the characterization given in Section 1, i.e. they pursue a basically linguistic interest but deal with pairs of languages that are ‘socio-culturally linked’. In fact, the majority of articles published in the journal *Languages in Contrast*, which was launched by the John Benjamins Publishing Company in 1998, deals with European languages, esp. Germanic and Romance ones. As far as the topics investigated are concerned, there is a preponderance of discourse-related studies, which may be due to the corpus-based methodology applied in most cases.

3 Establishing comparability

Just like linguistic typology, contrastive linguistics has to face the problem of “comparability of incommensurable systems” (Haspelmath 2008). In non-universalist frameworks (such as early structuralist linguistics and its modern successors), linguistic categories are only defined relative to the system that they form part of. Accordingly, the question arises whether categories from different linguistic systems can be compared at all, and if so, how such a comparison can be carried out. In very general terms, comparison can be defined as the identification of similarities and differences between two or more categories along a specific (set of) dimension(s). The categories compared must be of the same type, i.e. there has to be a set of properties that they have in common, or a superordinate category containing them. One major challenge for comparative linguistics thus is to determine the nature of that superordinate category (‘CS’) for any pair of categories under comparison:
In linguistic typology, the problem of “comparability of incommensurable systems” has been tackled in various ways. Haspelmath (2008) has argued that cross-linguistic comparison needs to be based on “comparative concepts”, i.e. analytic notions that are used to describe specific aspects of linguistic systems, e.g. ‘subject’, ‘case’, ‘(past/present/future) tense’, etc. For instance, a ‘subject’ in German does not have precisely the (system-internal) properties of a ‘subject’ in English. Still, ‘subject’ can be used as a comparative concept, in the sense of ‘grammaticalized neutralization over specific types of semantic roles’. Determining the extent of similarity as well as the differences between the instantiations of the comparative concept ‘subject’ in the languages under comparison is precisely the task that a relevant contrastive study has to carry out (cf. Rohdenburg 1974, König & Gast 2008: Ch. 6).

In contrastive linguistics, the ‘assumption of comparability’ for specific pairs of categories is reflected in, and supported by, linguistic output. Remember that contrastive linguistics has been defined as dealing with pairs (or groups) of languages that are socio-culturally linked, i.e. languages for which a substantial amount of bilingual output is available, for instance in the form of translations and parallel corpora. As Johansson (2000: 5) puts it, “[t]he use of multilingual corpora, with a variety of texts and a range of translators represented, increases the validity and reliability of the comparison. It can indeed be regarded as the systematic exploitation of the bilingual intuition of translators ….” The ‘hypothesis of inter-lingual commensurability’ is thus not a heuristic move but a fact of life reflected in the language of (balanced and fully proficient) bilingual speakers.

Bilingual output is also relevant to the question of (non-)equivalence between categories from different languages in another respect: Second language learners often identify categories from their L2 with categories from their L1 (‘inter-lingual identification’, ‘interference’, cf. Weinreich 1953). In other words, second language learners make an assumption of ‘inter-lingual equivalence’ that gives rise to non-target-like structures in their L2. In these cases, the (non-)equivalence of categories from different languages is not a question of heuristics but part of the object of study.

4 Comparison based on form: Consonant inventories

A phonological and morphophonological comparison of two languages is purely form-based insofar as it does not make reference to meaning or function. Specific aspects of phonological organization have figured prominently in (especially early) contrastive studies (e.g. Lado 1957). Given that phonemes are relational entities that can only be defined with reference to the system that they form part of and in fact constitute, they cannot easily be compared across languages. Let us consider the consonant inventories of English and German for illustration. A framework of comparison is provided by a classical structuralist analysis which is based on articulatory features of typical allophones instantiating the relevant phonemes (‘place of articulation’, ‘manner of articulation’ and ‘voicing’). Both the English phoneme /t/ and the German one /t/ can thus be regarded as instantiating the comparative concept ‘voiceless alveolar plosive’.
There are two basic types of relationships between such pairs of consonants: **near equivalence** and **non-equivalence**. The latter relationship is uninteresting in most cases – as the majority of pairs of consonants are obviously non-equivalent, say Engl. /p/ and Germ. /k/ – but there is a special case of non-equivalence that is highly relevant to contrastive studies, i.e. **partial equivalence**.

In the case of near equivalence two phonemes have a similar distribution and (in most contexts) similar phonetic realizations. For instance, the final consonant in the English word *bin* is both phonetically and phonologically similar to the final consonant in (the first person singular form of the German copula) *bin*, and the relevant phonemes have a similar distribution. The relationship between these phonemes is one of near equivalence (rather than ‘full equivalence’) because phonemes (as well as linguistic categories in general) are defined only relative to linguistic systems. This means that phonemes from different linguistic systems can never be fully equivalent.

A relationship of partial equivalence obtains when two phonemes are phonetically and distributionally similar but not (near) equivalent. For instance, the alveolar lateral of English and its German counterpart have a similar distribution but (partially) different phonetic realizations, as Engl. /l/, unlike German /l/, is velarized in a syllable-final position. If phonemes are regarded as sets of allophones, Engl. /l/ and Germ. /l/ can be said to overlap but not to be co-extensive:

\[
\begin{align*}
\text{(3) \quad \text{comparative concept}} \\
\text{‘voiced alveolar lateral’} \\
\text{Engl. /l/} \quad & \quad \text{Germ. /l/} \\
= \{l, \breve{l}, \check{l}\} \quad & \quad = \{l, \check{l}\}
\end{align*}
\]

As this example illustrates, the difference between near equivalence and partial equivalence is a gradual one. Partial equivalence can be assumed when the inter-lingual identification of two categories leads to considerable deviations from the target system in one of the languages involved. If German speakers identify the German /n/-phoneme with the English one, this will not lead to any noticeable deviation from native English phonology; the two categories are thus near equivalent. If, however, the /l/-phoneme of English is identified with the one of German, the pronunciation will be non-target-like in specific contexts (e.g. *[fɪl] instead of *[fɪl]*)]. Such ‘erroneous’ inter-lingual identification of categories from different languages leads to interference. The relationship between the categories involved can be called **pseudo-equivalence**; it holds between a pair of categories as conceived by an (unbalanced) bilingual speaker.

The contrastive method outlined above is illustrated in Diagram 2, where the role of bilingual data is taken into account. Each language is first analyzed in its own terms (e.g. by identifying alveolar laterals in English and German), and the ‘raw data’ is subject to a ‘preliminary comparison’ (e.g. by comparing Engl. *bin* to Germ. *bin* and Engl. *tall* to Germ. *toll*). Comparability is established on the basis of comparative concepts (‘voiced alveolar lateral’), and the pairs of categories thus identified are subject to a contrastive analysis against the background of bilingual output (e.g. the pronunciation of German L2-speakers of English).
5 \textbf{Comparison based on form and function: Temporal categories}

Most parameters of comparison investigated in contrastive studies are not purely formal but concern the mapping between form and function. As is well known from typological studies, this mapping is typically (and perhaps universally) many-to-many, i.e. each ontological category can be expressed using various linguistic categories, and each linguistic category covers a certain range of functions. This many-to-many relationship between form and function is illustrated in Table 1.

\begin{table}
\begin{tabular}{|c|c|}
\hline
Ontological categories & Linguistic categories \\
\hline\hline
OC_1 & LC_1 \\
OC_2 & LC_2 \\
OC_3 & LC_3 \\
OC_4 & LC_4 \\
\cdots & \cdots \\
\hline
\end{tabular}
\caption{Mapping from function to form and vice versa}
\end{table}

Still, the mapping from function to form is not entirely arbitrary. Roughly speaking, the domains of meaning covered by a given linguistic category must be semantically similar. In the ‘semantic map’ approach developed in linguistic typology (e.g. Haspelmath 1993, van der Auwera & Plungian 1998), semantic similarity is represented as proximity in an $n$-dimensional space. Such cross-linguistic models of form-function mapping can serve as a comparative concept in contrastive analyses. This will be illustrated with the example of tense categories in English and German.

As is generally the case in comparative studies, a certain amount of simplification is needed in the establishment of inter-lingual comparability. (4) represents a simplified model of temporal
reference that highlights those distinctions which are central to a comparison of English and German (cf. Declerck 2006):

(4)  
PAST  PRE-PRESENT  t₀  POST-PRESENT

Tense categories will usually cover contingent domains on a time axis as shown in (4). The ‘time spheres’ (roughly) covered by English and German tense categories are indicated in (5):

(5)  
PAST  PRE-PRESENT  t₀  POST-PRESENT

<table>
<thead>
<tr>
<th>English</th>
<th>German</th>
</tr>
</thead>
<tbody>
<tr>
<td>Past Tense</td>
<td><em>Präteritum</em></td>
</tr>
<tr>
<td>Present Perfect</td>
<td><em>Perfekt</em></td>
</tr>
<tr>
<td>Present Tense</td>
<td><em>Präsens</em></td>
</tr>
<tr>
<td>Future (will)</td>
<td><em>Futur (werden)</em></td>
</tr>
</tbody>
</table>

The same types of (non-)equivalence relations that were pointed out in Section 4 can be observed in (5). The relationship of near equivalence holds between the English *will*-future and the German *werden*-future. Again, the two categories cannot be said to be ‘fully equivalent’. Even though both future tenses are used for the future time sphere only (disregarding non-temporal uses like the epistemic ones), their distribution in English and German differs considerably. The reason is that there are different contrasting and competing tense categories in each language: The English *will*-future competes with the *going to*-future whereas the German *werden*-Futur competes with the *Präsens*. One consequence of this difference is that the English *will*-future is considerably more frequent than the German *werden*-Futur (in the spoken language), as the former, but not the latter category represents the ‘unmarked’ (semantically more general) choice vis-à-vis its primary competitor.

The English and German tense systems provide several examples of partial equivalence. For instance, the English Simple Present and the German *Präsens* are equivalent in some contexts but not in others. Roughly speaking, the contexts where the English Simple Present is used constitutes a subset of the contexts where the German *Präsens* is used. This, again, often gives rise to interference in bilingual speech (‘pseudo-equivalence’), e.g. when speakers of German use the English Simple Present for future time reference as a result of ‘inter-lingual identification’ of the two categories.

6 Comparison across functional domains

As was shown in the previous section, the comparison of categories associated with specific functions (‘tense’) typically departs from an ontological category (‘temporal reference’). In specific cases several types of ontological categories (as well as their manifestations in different languages) can be described in terms of the same comparative concept. A relevant example is provided by the two phenomena of relative clause formation and Wh-question formation in English and German (cf. Hawkins 1986). From an ontological (functional) point of view, these operations must be kept apart: Relative clauses are typically used to attribute a
property to some referent with the objective of enabling the hearer to identify that referent (the man who lives next door). By contrast, Wh-questions (Who will you marry?) elicit the value of a specific variable in an open proposition (‘You will marry x; who is x?’). However, in English and German both operations can be described in terms of the same comparative concept, i.e. ‘movement’ or ‘extraction’. Assuming that observable syntactic structures are (either in specific cases or always) the result of movement operations (one of the basic credos of pre-Minimalist Generative Grammar), both Wh-question formation and relative clause formation can be regarded as instances of extraction, differing only in terms of the (external) distribution of the relevant clauses. For instance, in (6) who is assumed to have been moved from its ‘base position’ (t) to its ‘surface position’:

(6) a. The man [who, you talked to t₁] is my brother.
   b. Who, did you talk to t₁?

The comparative concept of ‘extraction’ allows us to formulate generalizations across functional domains (relative clauses and Wh-formation). As Hawkins (1986) has shown, the operation of extraction is subject to different restrictions in English and German: English allows extractions out of finite complement clauses and nonfinite adverbial clauses, though not out of finite adverbial clauses. By contrast, German does not allow extractions out of finite or adverbial clauses at all (i.e. extractions are only possible out of non-finite complement clauses). This is shown in (7) and (8) and illustrated in Table 2 (from König & Gast 2008: 195).

(7) a. Who, did Charles think [that he saw t₁ in our garden]?
   b. The man [who, Charles thought [that he saw t₁ in our garden]] was my brother.

(8) a. *Wen, glaubte Karl, [dass er in unserem Garten t₁ sah]?
   b. *Der Mann, [den, Karl glaubte, [dass er in unserem Garten t₁ sah]], war mein Bruder.

<table>
<thead>
<tr>
<th>movement...</th>
<th>within clauses</th>
<th>across clauses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>complement clauses</td>
<td>adverbial clauses</td>
</tr>
<tr>
<td></td>
<td>non-finite</td>
<td>finite</td>
</tr>
<tr>
<td>German</td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Extractions in English and German

The example discussed in this section illustrates the role of linguistic theories in contrastive linguistics. They provide concepts and terms that can serve as a framework of comparison. Note that the ‘usefulness’ of any given model (for a given question) should be evaluated against the background of the (contrastive) generalizations that it allows one to make. Contrastive linguistics is thus often deliberately eclectic with respect to the theories and methods that it uses, and the choice of linguistic model may differ from one functional domain to another.

7 The use of corpora in contrastive linguistics

As the preceding discussion has shown, bilingual output plays an important role in contrastive linguistics in at least two respects: First, it provides a basis of comparison, or at least justifies the assumption of comparability; second, it constitutes the material on which contrastive generalizations are based. The existence of bilingual output has therefore been pointed out as
a central feature of contrastive linguistics, not least because it distinguishes this discipline from other types of comparative studies, especially typological ones.

Two major types of bilingual output can be distinguished: (i) data sets which instantiate each of the linguistic systems in ways that do not differ substantially from output produced by native speakers of the relevant languages (‘balanced bilingual output’); and (ii) data sets which are characterized by deviance from relevant output produced by native speakers in one of the languages involved (‘unbalanced bilingual output’). Balanced bilingual output is represented by (high quality) translations and parallel corpora based on such translations. Unbalanced bilingual output is represented by the non-target-like language of second language learners. Such data has also been collected in large samples of texts in the form of ‘learner corpora’. Each type of resource can be used for different purposes.

As a general tendency, parallel corpora are associated with quantitatively oriented (often distributional) studies of specific linguistic features in discourse. The results obtained in such studies are often relevant to translation studies. In recent years, parallel corpora have played a particularly prominent role in contrastive linguistics based in Scandinavian countries (e.g. Aijmer et al. 1996, Johansson 1998b). It is also in this research context that extensive parallel corpora have been compiled, e.g. the English-Norwegian Parallel Corpus, which was assembled between 1994 and 1997 at the University of Oslo. This corpus contains (pairs of) texts that have been translated in both directions, i.e. there are English originals with Norwegian translations and vice versa. Such ‘bidirectional’ corpora allow for the investigation of rather subtle questions concerning the theory and practice of translation, e.g. ‘hidden’ interference phenomena and translation norms.

While parallel corpora provide (balanced) bilingual output, learner corpora are ‘bilingual’ in a different way: they contain only data from one language, which is, however, produced by second language learners and consequently exhibits features of the leaner’s L1. One of the most comprehensive learner corpora available – the International Corpus of Learner English (ICLE) – has been compiled at the University of Leuven under the coordination of S. Granger (cf. Granger 1998). It contains more than 3 million words produced by native speakers of more than twenty different languages. Even though the computerized analysis of interlanguage need not pursue a didactic purpose, it obviously lends itself to several pedagogical applications and has in fact become a central component of technology-enhanced learning in recent years.

References


