The corpus experience at INESC

The INESC Natural Language Group has done research based on or aimed at corpora since 1987:

- using a test corpus for evaluation of an MT system  
  (Santos, 1988)
- compiling a "sentence" corpus for driving a computational grammar  
  (Santos, ed., 1992)
- developing a set of corpora tools for grammatical research  
  - corpus browser  
  - some statistical measures  
  (Medeiros, 1992; 1995a)
- obtaining some quantitative measures about Portuguese  
  - part-of-speech homography  
  - "lexical filling"  
  (Medeiros et al., 1992; Santos, 1994)
- syntactic behavior of one verb  
  (Bacelar do Nascimento et al., 1992)
- using specialized error corpora for evaluating existing spelling checkers  
  (Medeiros, 1995a; Medeiros, 1995b)
- producing corpus-based research on contrastive matters  
  - on the two variants of Portuguese (European versus Brazilian)  
    (Wittmann & Pereira, 1994; Barreiro et al., 1995; Santos et al., 1995)  
  - on the tense and aspect systems of English and Portuguese  
    (Santos, 1994a; 1994b; 1995a)
Why do contrastive analysis based on parallel corpora

Contrastive analysis is generally presented with the following goals:
- the wish to know more about the two languages
- the wish to know more about natural language in general
- to help translation
- to help foreign language teaching

In my case, it stems from the need to
- understand Portuguese
- while using what had already been written on tense and aspect in English, without neither assuming identity nor difference

Parallel corpora are data for contrastive analysis which reflect performance. A competence-oriented contrastive analysis has the danger of concentrating on possible, but not preferable, ways of expressing other language's categories.

Languages differ mostly not in what they can say, but in what they actually say.

This is not to say that matters of competence are alien to contrastive studies of this sort: Corpus-based studies rely primarily in unbiased data, but secondarily on competence from the part of the analyst.
Parallel corpora and translation

Even though it is undeniable that the only objective data for a fine-grained contrast between two languages are translated texts, given that the translators try to "say the same thing" in the other language, there is a host of problems with the use of translations as data:

1. It is known that the translation is not always (if ever) an equivalence relation.

2. Because of the influence of the source language, the translator must create a text which cannot be free from "translationese".
   (the influence of the source language may be involuntary, or may be caused by the attempt to struggle it, "antitranslationese")
   Anyway, a translated text is never a good exemplar of target language.

3. Languages express a maximum of compatible meanings/connotations through the same marker, i.e., a good context of use of a grammatical device is one in which many reasons can be adduced for its use.
   However, the set of reasons is a function of the whole grammatical system, which makes it highly improbable that two markers in two different languages will correspond to the same set.
   In other words, there is loss - and there may be addition of unwanted connotations as well.
A fundamental issue

More importantly, though, is: how to compare two languages without an independent term of comparison?

"It is logically impossible to engage in contrastive analysis without postulating common categories of one sort or another since, more generally, it is logically impossible to compare any two entities without using the same frame of reference" (van Buren, 1990:85).

On the other hand, there is no guarantee that underlying semantic categories in any two languages cover the same range or are involved for the same situations.

Instead of expecting a language-independent set of categories, one can "measure" one language via the other language's categories. This is, in fact, what people has always (implicitly) done. In general, using English as the measuring rod (as noted in Santos, 1992).

For example, according to Smith's (1991) parametric approach, there is in English a perfect match between the perfective viewpoint and situation type, contrarily to all other languages she analyses. (!)
Or, following Dorr (1991) in using LCS as an interlingual structure, and parameterizing divergences between languages, it is obvious that LCS is closest to English.

This is, also, what translation does: it expresses one language's categories through those of the other (the target language).

One can thus state that A of L1 transmits what B of L2 does,
    that C of L1 is not expressible in L2
or that the distinction between D and E in L1 has no grammatical counterpart in L2
On translation equivalence

If two features are translationally related, do they mean the same?

At least three alternative answers:

1) No, they are the closest way in each language to say the particular thing meant

2) No, but they have maximal overlapping between what they say

3) No, but they are the case where more (different) things are kept said.

My answer is: they are instances of categories of different languages, and therefore secondary or derived implications in one language may constitute the core of the translated category in the other language.

Example: English marks "in progress" and encodes lexically "has a inherent end". On the other hand, Portuguese has a category "temporary state". It is natural (and it actually happens frequently) that something in progress in English is translated by something temporary (after all, something which has an inherent end and is in progress is bound to finish). However, a Portuguese temporary state is not necessarily something which is in progress and ends.

Example 2: Portuguese marks "according to plan", English marks present relevance. Often, the occurrence of what is accorded to plan is relevant to the present, and thus these categories are translationally related.
Basic assumptions

1. The analysis of each language must be done following the categories proposed by the language itself. Therefore, a different metalanguage should be used (which is trivially simple if one uses a natural metalanguage).

2. Languages should be contrasted without resorting to a third metalanguage (interlingua, independent KR language, or the like), rather, using each other's metalanguage.

3. Objective data for this contrast are furnished by an analysis of translations in the two directions. This analysis must be fine grained, though, because

   a) translation is not unique:
      - there is choice among the several features rendered by the source utterance
      - there is information which is conveyed unintentionally (it is not functionally relevant)

   b) translation is not perfect
      - the translator may have a deficient knowledge of the source language
      - s/he may be influenced by formal similarity (interference)
Consequences for contrastive studies

1. The contrast must always be directional, i.e.,

   L1 seen through L2, and L2 seen through L1.

2. Some precautions concerning translations as data must be taken, since

   a) the language of a translation is not a good representative of the target
      language: it will privilege the categories with which it translates those of
      the source language and will have less emphasis on its own categories which
      are not shared by (or related to) the source language.

      Example: Use of (unambiguous) habitual Imperfeito in translated English
      decreases drastically

   b) there will be arbitrary choices for translation, which should not be taken as
      representative of either the source or the target language

      Example: choices between inception or perception; choices between
      unattended perception or physical ability.

   c) there will be cases in which the lack of parallelism is so huge that one
      cannot compare (formal) categories

      Example: different sentence structure; radically different lexical items.
A model for contrastive studies based on translation: the Translation Network

It contains two aspectual networks (inspired by Moens, 1987), where categories and operators are strictly monolingual. Then, a set of unidirectional arcs links the source language to the target language, modelling translation.

This model was conceived to formalize the temporal and aspectual categories in two languages, and the comparison between them and, therefore, is a competence model (of language and of translation).

But it was also designed to describe/formalize real translations, and, therefore, to be integrated in a translation browser that evaluate the quality of a translation (according to the modelled parameters).

Examples:
A model for contrastive studies based on translation: the Translation Network

It contains two aspectual networks (inspired by Moens, 1987), where categories and operators are strictly monolingual. Then, a set of unidirectional arcs links the source language to the target language, modelling translation.

This model was conceived to formalize the temporal and aspectual categories in two languages, and the comparison between them and, therefore, is a competence model (of language and of translation).

But it was also designed to describe/formalize real translations, and, therefore, to be integrated in a translation browser that evaluate the quality of a translation (according to the modelled parameters).

Examples:

1. L₁ faz uma distinção a nível gramatical; L₂ a nível lexical
2. A é opcional em L₁, mas a distinção é obrigatória em L₂
3. L₂ é vaga

Diana Santos 8 XI Congresso da APL, 1995