The Routledge Handbook of Audiovisual Translation

Luis Pérez-González

Corpus-based audiovisual translation studies

Publication details
Maria Pavesi
Published online on: 06 Sep 2018

How to cite :- Maria Pavesi. 06 Sep 2018, Corpus-based audiovisual translation studies from: The Routledge Handbook of Audiovisual Translation Routledge
Accessed on: 08 Jun 2023

PLEASE SCROLL DOWN FOR DOCUMENT

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Corpus-based audiovisual translation studies

Ample room for development

Maria Pavesi

A history of the area

This chapter describes the development and use of corpora within audiovisual translation (AVT), exploring the methodological, theoretical and descriptive insights that corpora have originated in the field. In recent years researchers have become increasingly aware that large and principled electronic collections of audiovisual dialogue and computer-assisted methods of analysis are needed to carry out empirically validated investigations of AVT (e.g. Heiss and Soffritti 2008, Freddi and Pavesi 2009a, Baños et al. 2013, Pavesi et al. 2014). Corpus-based articles, monographs and doctoral theses have proliferated, drawing on both lower-scope data collections and larger databases that bring corpus-based research on AVT into line with other areas of Corpus-based Translation Studies (CBTS). This general approach to research on translation can be placed at the intersection between the theoretical and methodological frameworks of Descriptive Translation Studies and Corpus Linguistics since it responds to the call to look for distinctive features, patterns, norms and universals in translated texts by relying on extensive authentic data. Launched at the beginning of the 1990s by Baker’s seminal work (Baker 1993, 1996), CBTS embody an empirical, intersubjective as well as descriptive, rather than prescriptive, viewpoint on translation as developed in Toury’s (1995/2012) target-oriented polysystemic theory (Laviosa 2011).

There are sound methodological, theoretical and applied reasons to create and exploit corpora of AVT. The first one derives from the need to provide reliable generalizations and achieve descriptive adequacy by moving beyond the limited scope of single case studies, while overcoming the failures and inconsistencies of introspection and intuition. Corpora therefore allow the researcher to reveal cumulative effects that would not emerge through the examination of individual translations. These incremental effects, arising from the repetition of the same patterns across texts, account for the uniqueness of AVT registers by pinpointing their typicalities. They can also unveil evaluative and ideological meanings that prevail in different cultures and in different translation types (cf. Baker 2006: 13). In addition, corpus analysis can highlight areas that are recurrently problematic while developing translators’ awareness about ‘hidden’ linguistic and stylistic regularities in translated
and domestic screen dialogue. In this way, corpus analysis can ultimately be exploited to improve the translation quality of audiovisual products (see Freddi 2012: 382).

Two main foci of attention have been observed in corpus-based audiovisual translation studies (CBAVTS) since the beginning: (i) the description and explanation of the relationships between translated texts and original/non-translated texts and (ii) the description and explanation of the relationships between source texts and target texts. CBAVTS thus adopt both a target-oriented and source-oriented comparative approach, with investigations typically looking for similarities and contrasts within and between corpora and sub-corpora. Different emphases, however, have characterized research on the different modalities of dubbing, subtitling and audio description (AD), as will be shown in the remainder of this chapter.

Analysts in the field often work with translation-driven corpora compiled or assembled for a specific purpose or to investigate a specific class of phenomena, such as discourse markers, cultural references and pragmatic routines. More rarely, AVT corpora are multi-purposed, self-standing, flexible resources suitable to address a wide range of research questions. In most cases, these databases are monomodal written or spoken corpora, the latter comprising transcriptions of spoken dialogues at different levels of granularity and detail that can be further enriched with linguistically encoded paralinguistic and nonverbal information. As a result, spoken corpora crucially differ from speech corpora or multimodal corpora, in which the audio alone or in combination with the visual is also available (Adolphs and Carter 2013). In terms of division of labour, whereas spoken corpora inevitably present a partial picture of the dialogic exchange, they provide a good basis for large-scale investigations leading up to quantitative generalizations about the linguistic configuration of audiovisual dialogues. By contrast, speech and multimodal corpora are most appropriate for in-depth qualitative analyses centring on the interactions between the various signifying codes co-deployed on screen (Heiss and Soffritti 2008, Valentini 2008, Soffritti this volume).

Main issues in corpus-based audiovisual translation studies

Corpus compilation: types of corpora and construction criteria

Corpora come in different sizes but in AVT, due to the difficulties and economic costs of creating spoken corpora, they tend to be small do-it-yourself collections. Following Zanettin (2012), we can distinguish three main types of translation-driven corpora: monolingual comparable corpora, bilingual comparable corpora and parallel corpora. Monolingual comparable corpora are made up of two sets of texts in the same language, chosen using similar design criteria. Whereas the first set contains translations into a target language, and for this reason is also called translational corpus, the second set contains comparable domestic texts. By contrast, bilingual comparable corpora bring together non-translated texts belonging to different languages but sharing similar conceptual domains as well as discoursal and pragmatic functions. Lastly, parallel corpora include a principled set of source language texts together with their translations into a given target language. Monolingual comparable corpora are suitable to study the unique features of translated vis-à-vis non-translated varieties of the same language, while with parallel corpora researchers can address issues such as functional equivalence, translation shifts and translation strategies. Bilingual comparable corpora, in turn, allow the researcher to carry out preliminary contrastive analyses (Toury 1995/2012). At best, corpus-based research draws on a combination of corpus resources and uses both comparable and parallel components to formulate more accurate and reliable hypotheses about translated
language (Saldanha and O’Brien 2013: 69). In CBAVT, researchers also resort to reference corpora, i.e. representative corpora of the source, target or translated language, functioning as baselines for comparison.

CBTS have identified various methodological issues in corpus building whose relevance extends to corpora of AVT. Such issues include representativeness, size, comparability, transcription conventions, annotation and corpus alignment. When constructing a corpus, audiovisual products should be chosen to be as representative as possible of the population at large or of specific genres or product types (e.g. audio described films, subtitled documentaries, TV series for adolescents). Related to representativeness are the issues of size, selection criteria and sampling, balance and homogeneity. In general, larger corpora have a better chance of being representative of the target population, although this advantage tends to tail off when the corpus exceeds a certain size. Moreover, whereas big corpora enable the scrutiny of a wide range of phenomena, including infrequent vocabulary items and spoken pragmatic features, they may generate unmanageable amounts of data that call for laborious sorting and further sampling. A smaller, specialized corpus may hence be preferred if we move from restricted and well-formulated hypotheses. It should be finally pointed out that corpora compiled to be general corpora need to be much larger and entail more careful design criteria than specialized corpora (Zanettin 2012: 41), as most of those used in AVT research.

Regardless of size, clear selection and sampling criteria are required to make sure that the corpus is representative and balanced. This means that it should contain an adequate coverage of the products under investigation without being skewed in favour of any specific generic or textual category. Uniformity is also required and items should be included that belong to the same category that researchers aim to investigate.

In CBAVT research, another important criterion is given by comparability between the different components of the corpus or, to a different extent, between the corpus and the reference corpora used. Comparability should be weighed up vis-à-vis factors including completeness of the audiovisual texts and date of release, but also composition and genre, so as to obtain corpus components that are modelled one over the other (Olohan 2004: 42, Zanettin 2012: 48). But there is also a trade-off between comparability and representativeness as different countries privilege different audiovisual genres in both films and TV series, often varying in terms of type and amount of audiovisual productions.

Specific to spoken corpora, and hence to many AVT corpora, is the issue of transcription. As a graphic representation of talk, transcription first requires the researcher to interpret data by considering the prosodic, gestural, textual, and environmental elements that interact with speech and constitute communicative exchanges in a multimodal context. The process, however, forcibly entails a selection and a reduction of the data that will be visualized graphically in the transcript (Adolphs and Carter 2013). Based on the aims of the project, decisions thus have to be made on how detailed the transcript can or has to be (Bonsignori 2009), although to allow the smooth electronic processing of the data, spoken corpora in AVT typically use some form of standardized orthographic transcription. For this reason, most sociolinguistic variation is necessarily levelled out, and can only be re-introduced through some form of annotation. Despite the centrality of transcription in spoken corpora, few published AVT studies explicitly state the rules and the procedures followed in managing spoken data.

Finally, as copyright clearance must be required for both original and translated audiovisual products, copyright issues are a major stumbling block in corpus building, slowing down the process or even halting it, and seriously limiting accessibility. In addition, these issues can easily hamper the representativeness of the corpus if the final compilation of audiovisual products relies on the clearances obtained.
Three corpora of audiovisual translation

Not many AVT publications offer a fully fledged description of how the corpora used for analysis were built and even fewer, if any, make it explicit whether clearances from copyright holders were received. However, a few corpora are available or are being created that can exemplify the methodological choices carried out before and during corpus compilation. The TIWO (Television in Words) Audio Description Corpus (Salway 2007) was built to investigate the language of AD, starting from the hypothesis that this special language is characterized by the prevalence of features directly related to its communicative functions. More precisely, the analysis to be carried out through a corpus-based methodology was intended to ‘identify and describe a special language in terms of statistically significant differences between linguistic features in a corpus of [audio description] scripts and a general language sample’ (Salway 2007: 154). The TIWO is thus implicitly a monolingual comparable corpus, whose internal component is made up of 91 British English AD film scripts, totalling 618,859 words. To obtain a corpus representative of the translation modality, scripts were included that covered different ways of carrying out AD in Britain: action, children’s animation, children’s live action, comedy, dark, period drama, romantic, thriller and miscellaneous. The TIWO Audio Description Corpus includes complete scripts directly obtained from three major producers of AD, a fact that pre-empts copyright issues.

The Pavia Corpus of Film Dialogue (PCFD) is a parallel and comparable spoken corpus built to investigate original audiovisual dialogue and its dubbing translation (Pavesi 2014; PCFD, n.d.). The parallel component is made up of 24 orthographically transcribed American and British film dialogues and their dubbed Italian translations (about 500,000 word tokens). This component is unavoidably unidirectional owing to the translation policies pursued in Anglophone countries (which subtitle rather than dub foreign audiovisual products). The corpus also currently contains a comparable component of 24 Italian original films (about 220,000 word tokens). Constructed to research the representation of spokenness in film language and AVT, the PCFD alone and in combination with reference corpora of English and Italian can be used to address a wide array of sociolinguistic and pragmatic issues so as to eventually delineate a profile of contemporary dubbed Italian. The films to be included in the corpus were therefore chosen to be representative of mimetic audiovisual products, that is products that portray naturalistic language use, in compliance with a series of sampling criteria (Pavesi 2014: 37). Films had to:

(i) be set in contemporary times, representing contemporary dialogue;
(ii) present a prevalence of situations that in real life elicit free and bidirectional spontaneous spoken language;
(iii) have been released at regular intervals within a fixed time span (from 1996 to 2009);
(iv) have been successful, both with the critics and the general public.

The broad selection of films enhanced the representativeness of the corpus with the aim of offering a wide range of represented settings, situations, topics and characters, and generating variation along the portrayed sociolinguistic parameters. However, due to copyright issues, a compromise had to be reached between representativeness and availability, which meant that a few previously selected films had to be excluded from the final sample (Pavesi 2014).

As for alignment and systematic interrogation, the PCFD was converted into a relational database (Freddi and Pavesi 2009b, Freddi 2013), which permits users to carry out queries.
beginning from either the original or the translated components. Several parameters were added to the database: textual and contextual variables, i.e. character speaking, scene type and linguistic event (e.g. on the phone, on television), together with individual variables including accents (e.g. French accent, Spanish accent), accompanying paralinguistic behaviour (e.g. whispering, shouting, giggling), and salient non-linguistic behaviour (e.g. waving, kissing). Table 20.1 shows some of the parameters instantiated during the search of ‘you know?’ in the database.

The biggest monomodal corpus envisaged so far within a corpus-based approach to AVT is CORSUBIL (Corpus de Subtítulos Bilingües Inglés-Español): a bilingual corpus made up of English and Spanish subtitles, both meant for mainstream viewers, extracted from significant American movies, aiming at 18 million words, 9 million for each sub-corpus (Rica Peromingo 2014, Rica Peromingo et al. 2014). The considerable dimensions of the corpus are needed to address the wide goals of the project, both descriptive and applied, as the corpus is intended to provide data relevant for bilingual lexicology and lexicography, language teaching and language learning. Since among the functions of film dialogues we find the simulation of orality, one specific aim of research on CORSUBIL is the creation of bilingual lists of conversational routines in English and Spanish, allowing the analysis, for example, of polite speech-act formulas and discourse markers in translation and cross-linguistically. Differently from smaller corpora, the representativeness of the corpus is warranted by the very broad and principled selection of the most important movies in American film history, from the beginning of the talking movies onwards up to the twenty-first century. The equal division into five periods will permit researchers to have a general overview of the language of American film subtitles and the corresponding Spanish translations together with a diachronic and more time-specific perspective on audiovisual dialogue.

Tools for corpus analysis

The most common tools for corpus analysis in AVT research are frequency counts, monolingual concordances and bilingual or aligned concordances. Both individual words and clusters of words can be searched and frequency counts may be compared across corpora or sub-corpora to examine the differences between the source and the target texts or between the translated texts and comparable texts in the same target language (Freddi 2009, 2012; Romero-Fresco 2009, 2012). To test the hypothesis about the specificity of the language of AD, Salway (2007: 158–159), for instance, calculated the ratio between the relative frequency of words in the TIWO corpus and the relative frequency of the same words in a sample of the British National Corpus, thus obtaining a list of items that typify AD. In turn, Bonsignori et al. (2012) compiled frequency lists of greeting and leave-taking formulas in a DIY corpus of Anglophone films and their Italian dubbing translations both to compare the levels of formality and naturalness in source and target texts and infer the motivations that underlie recurrent translation choices. In both studies, frequencies pertained to the words contained in raw (un-annotated) corpora. Like other corpora, however, AVT databases can also be enriched with tags that allow the computing of specific phenomena. In Pavesi’s (2013) investigation of demonstratives in the PCFD, all 1,678 English pronouns were manually annotated for syntactic role, pragmatic function and translation strategy into Italian. The annotated texts were then searched with the program AntConc 3.2.1w to gather information on the frequency of the various translation strategies, which was found to depend on the subject or object function of the demonstrative
Table 20.1 Extract from the Pavia Corpus of Film Dialogue displaying various types of data.

<table>
<thead>
<tr>
<th>Title</th>
<th>Cue</th>
<th>Character</th>
<th>Scene</th>
<th>Attitude</th>
<th>Text</th>
<th>Translations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lost in Translation</td>
<td>406</td>
<td>JOHN</td>
<td></td>
<td>kisses</td>
<td>Yeah, call those guys, you know and I - I'm gonna be back on Sunday and I - I love you, you know? Okay?</td>
<td>Si, chiamali, è una buona idea! E poi, io, io, io torno domenica e - e - e . . . ti amo, lo sai. Okay.</td>
</tr>
<tr>
<td>Match Point</td>
<td>407</td>
<td>NOLA</td>
<td></td>
<td>clears throat</td>
<td>Erm . . . I thought he was very handsome. You know? And I told you I was just . . . I was overwhelmed with attention. So what about you and Chloe?</td>
<td>Em . . . bè, lo trovavo molto attraente. Inoltre, te l'ho detto, sono stata . . . sopraffatta dalle sue attenzioni. E la cosa fra te e Chloe?</td>
</tr>
<tr>
<td>The Holiday</td>
<td>18</td>
<td>IRIS</td>
<td></td>
<td>sighs</td>
<td>I was head over heels, you know? Everyone knew.</td>
<td>Ero innamorata cotta, lo sai? Lo sapevano tutti.</td>
</tr>
<tr>
<td>Two Lovers</td>
<td>1166</td>
<td>REUBEN</td>
<td>pouring some wine</td>
<td>Hebrew accent</td>
<td>You know, we celebrate . . . we celebrate the New Year twice a year, you know? Because at Rosh Hashanah and . . . All right . . .</td>
<td>Si, noi festeggiamo . . . noi festeggiamo il nuovo anno due volte l’anno. La prima è Rosh Hashanah . . . e adesso. Ecco.</td>
</tr>
</tbody>
</table>
Table 20.2 Concordances of *amico* in the translational component of the Pavia Corpus of Film Dialogue.

<table>
<thead>
<tr>
<th>Concordance</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>sono rotto le palle! Ma che cazzo fai? Quello è <em>amico</em> mio. Ma perché non te ne vai?</td>
<td><em>amico</em> friend</td>
</tr>
<tr>
<td>egne saltate per settimane. Che significa, Eric? <em>Amico</em> mio, che ti succede? Oh, non è.</td>
<td></td>
</tr>
</tbody>
</table>

Table 20.3 Concordances of ‘smiles’ plus adverbs in –*ly* from the audio description of *The English Patient* (Palmer and Salway 2015: 137).

<table>
<thead>
<tr>
<th>Character</th>
<th>Concordance</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Katherine</td>
<td>smiles awkwardly.</td>
<td></td>
</tr>
<tr>
<td>She</td>
<td>smiles benignly.</td>
<td></td>
</tr>
<tr>
<td>Lady Hammond</td>
<td>smiles delightfully as Katherine crossed the courtyard.</td>
<td></td>
</tr>
<tr>
<td>Almasy</td>
<td>smiles faintly.</td>
<td></td>
</tr>
<tr>
<td>She</td>
<td>smiles fondly at the man with the melted face.</td>
<td></td>
</tr>
<tr>
<td>He</td>
<td>smiles ruefully and lowers his head, looking pensive.</td>
<td></td>
</tr>
</tbody>
</table>

and its deictic role in the dialogue. More recently, other frequency measurements have been used to investigate corpora of AVT, such as type/token ratios and lexical density (Formentelli 2014, see below).

Whereas frequency lists unveil overall patterns in audiovisual texts, monolingual concordances are suitable for in-depth analyses of word patterning. When extracting concordances in a corpus, all the lines containing the searched word, or node, are displayed together in a vertical list, hence providing in one go access to the various contexts in which lexical units and clusters occur. Table 20.2 displays the KWIC (Key Word In Context) concordances for *amico* ‘friend’, a recurrent lexical item in Italian dubbese where it often functions as a sociolinguistically marked vocative (Pavesi 2005: 50). The concordance lines are extracted from the PCFD.

The research potentials of concordancing are many. With a view to uncovering how much information AD conveys about characters’ thoughts as opposed to actions, Palmer and Salway (2015) inspected the concordances of individual items that had turned out to be very frequent in the specific discourse. Starting from –*ly* adverbs and their combinations with the forms ‘looks’ and ‘walks’ in the TIWO corpus, the analysts discovered that, although each action verb has a different set of preferred adverb collocates, most of these add information about mental states (Palmer and Salway 2015: 142–143). *Looks*, for instance, is shown to be followed by ‘directly’ (*f* = 9), ‘anxiously’ (*f* = 8), ‘sadly’, ‘steadily’, ‘thoughtfully’ (*f* = 7), ‘nervously’ (*f* = 6), ‘fearfully’, ‘grimly’, ‘longingly’, ‘quizzically’ (*f* = 4). By carrying out a closer analysis of the concordances of the high-frequency verb form ‘smiles’ (see Table 20.3), it became clear that several recurrent word combinations give insight into characters’ thoughts (Palmer and Salway 2015: 139).
With bilingual concordances the researcher can explore the correspondences for given expressions starting from either the source or the target texts in parallel corpora, in this way examining cross-linguistic correspondences and checking whether given translation shifts are systematic. Rica Peromingo (2014) compared phraseological units in source and target texts by examining the bilingual concordances in a small sample of CORSUBIL. These show that as a rule there are no direct or one-to-one translations for the English discourse markers transferred into Spanish. Rather, one-to-many correspondences obtain in both directions: the same English expression has various renderings in Spanish, whereas the same Spanish phraseological unit translates different English formulas, as illustrated in the bilingual, aligned concordances for ‘sorry’ and ¿Qué pasa? in Table 20.4.

Bilingual and monolingual concordances can in fact be combined when researchers draw on parallel and reference corpora simultaneously. In a study aimed to map out register shifts in racist discourse, at the outset Mouka et al. (2015) annotated for categories of appraisal a small corpus of Anglophone films subtitled into Greek and Spanish. The intensifying or neutralizing effect of translation choices was then assessed by inspecting the bilingual context of the negative expression. Such contextual interpretation was later checked against monolingual concordance lines and collocates for the same expression in reference corpora. In Example 1 below, the weakening of the negative evaluation implicit in ‘border jumpers’ is first indexed by the omission of the vulgar adjective in the Greek translation.

Table 20.4 Bilingual concordances for ‘sorry’ and ¿Qué pasa? from the CORSUBIL corpus (Rica Peromingo 2014: Tables 1 and 2).

<table>
<thead>
<tr>
<th>English subtitles</th>
<th>Spanish subtitles</th>
</tr>
</thead>
<tbody>
<tr>
<td>21 00:02:42,640→00:43,960 Sorry.</td>
<td>21 00:02:42,640→00:02:43,960 Perdón.</td>
</tr>
<tr>
<td>601 01:01:38,800→01:01:41,520 Sorry, but you will have to leave.</td>
<td>563 01:01:38,800→01:01:41,520 Lo siento. Tiene que irse.</td>
</tr>
<tr>
<td>33 00:06:52,454→00:06:54,331 I’m sorry. I’m sorry, sir.</td>
<td>31 00:06:52,454→00:06:54,331 Perdóneme, señor, perdóneme.</td>
</tr>
<tr>
<td>77 00:07:48,880→00:07:51,280 – What’s the matter? – it’s Michael</td>
<td>72 00:07:40,413→00:07:43,211 – ¿Qué pasa? – Michael no está.</td>
</tr>
<tr>
<td>513 00:52:34,200→00:52:25,680 What is it?</td>
<td>459 00:52:24,200→00:52:25,680 ¿Qué pasa?</td>
</tr>
<tr>
<td>886 01:37:45,280→01:37:46,280 What’s wrong?</td>
<td>806 01:41:27,568→01:41:29,661 ¿Qué pasa?</td>
</tr>
</tbody>
</table>
Example 1

[en] And now some fucking Korean owns it who fired these guys and is making a killing because he hired 40 fucking border jumpers.

[el] Τώρα το 'χει ένας Κορεάτης, που απέλυσε τους δικούς μας και θησαυρίζει επειδή προσέλαβε λαθρομετανάστες.

[Back translation] Now a Korean owns it who fired our guys and is making a killing because he hired illegal immigrants.

The analysis of the concordance lines for ‘border jumper’ in the English reference corpus enTenTen12 further unveiled that the expression in the source linguaculture is ‘used almost exclusively in a negative and highly disparaging sense (border jumpers want our wealth; drug smugglers, human traffickers, border jumpers and other assorted criminals; border jumpers are slapping those legal criminals)’ (Mouka et al. 2015: 58). This highly negative colouring is not shared by its subtitled translation λαθρομετανάστες ‘illegal immigrants’, a more neutral expression belonging to standard and authoritative Greek.

The patterns emerging from these research analyses are all relevant to the identification of systematic linguistic and translational behaviour in source, target and reference texts as will be illustrated in the following sections.

Approaches to corpus analysis

Quantitative and qualitative corpus analyses in AVT research can be carried out following both a semasiological (form-to-function) and an onomasiological (function-to-form) approach. By pursuing a semasiological approach, researchers start from individual words or sequences of words to uncover underlying patterns of meaning and regularities. By contrast, with an onomasiological approach, the analyst starts from a given meaning and looks for features, constructions and patterns that express that meaning. In her study of compliments in dubbed dialogues, Bruti (2009), for instance, has argued for a procedure that integrates the retrieval of pragmatically loaded words such as ‘nice’ with the onomasiological analysis of the speech acts identified by scanning the corpus manually. In addition corpora can be annotated by the analyst for functional categories such as speech acts, humour and metaphors by adding ‘special marks making the token “visible” to the corpus software’ (Mikhailov and Cooper 2016: 11) to ease retrieval, inspection and quantification. This is however a lengthy procedure, feasible only with small corpora. As a result, most quantitative research in CBAVTS has adopted a semasiological approach, whereby researchers formulate some specific hypotheses and check on the translational behaviour of previously identified grammatical, lexical, discoursal or pragmatic features.

Starting from a now well-established corpus linguistic approach, AVT researchers have an additional analytical procedure at their disposal. Evidence of variation and regularity patterns can be collected inductively as it emerges from the data without the constraints deriving from hypotheses formulated in advance. A successful application of such corpus-driven analysis is Freddi’s (2009, 2012) study of phraseology in both original and dubbed film language. By generating frequency lists with the software program Wordsmith Tools, the author first extracted the top word clusters in the English dialogues and then checked to which degree and with which functions the formulaicity detected in the original texts transferred to the Italian dubbing. High-frequency formulas like ‘Oh my God’
were rendered with the same formulas (in this case, *Oh mio Dio*, ‘Oh my God’) in translation, while other repeated clusters such as ‘you know what’ were dubbed with more varied and hence creative translations: *la sai una cosa, sai che ti dico, facciamo così*, all meaning ‘you know what’, together with *senti* (‘listen’) and *okay* (‘okay’). By combining a corpus-based and a corpus-driven approach, Forchini (2013) tackled the translation of English vocatives, a phenomenon that has proved to be problematic in Italian dubbing. The researcher first selected a group of frequent familiarizers in English conversation (‘man’, ‘guys’, ‘buddy’ and ‘dude’) to later retrieve the frequency of these lexical items, their collocations (i.e. recurrent juxtapositions of words) and cluster combinations in a small diachronic corpus of original Anglophone films dubbed into Italian. The combination of the various computer-assisted analyses contributed to the definition of a usage profile for each vocative while highlighting how translational norms in Italian dubbing have evolved in time towards pragmatic equivalence.

A close-up on applications in corpus-based audiovisual translation studies

**Naturalness and the register-specificity of audiovisual translation**

CBAVTS have focused on key research questions that are connected to the nature of audiovisual texts and the different AVT modalities. The issue of spokenness or naturalness is strictly pertinent to screen dialogue, which represents a type of ‘scripted/constructed dialogue’ (Bednarek 2010: 63) or language ‘written to be spoken as if not written’ (Gregory 1967: 191–192). It is no surprise, therefore, that the issue of naturalness has taken centre stage in many investigations of the language of dubbing and its alignment or disalignment with spontaneous conversation though the study of a wide array of linguistic phenomena. These include discourse markers, greetings and leave takings, general extenders, word clusters, intensifiers, interjections, hesitations, marked word orders, personal and demonstrative pronouns, questions, slang and vocatives. To address the issue of naturalness, the most reliable target-oriented approach involves the comparisons between dubbed language and spontaneous spoken language, on the one hand, and between dubbed dialogue and domestic dialogue, on the other hand. Starting from a definition of naturalness as ‘*nativelike selection of expression in a given context*’ (Romero-Fresco 2009: 51, italics in the original), Romero-Fresco (2009, 2012) displayed the full potentialities of such a comparative methodology by delving into the behaviour of a group of intensifiers (*en serio, de verdad* and *de veras*, ‘really’), discourse markers (*veamos, vamos a ver* and *a ver*, ‘let’s see’), and transition markers (*(muy) bien* and *bueno*, ‘so’/‘okay’/‘well’) in dubbed Spanish. In his investigations, the author made use of three different corpora:

1. a parallel corpus of transcripts of the American TV series *Friends* and their dubbed Spanish versions, totalling about 300,000 words;
2. a comparable corpus consisting of the dubbed *Friends* dialogues and the original, *Friends*-inspired Spanish sitcom *Siete Vidas*, totalling about 300,000 words;
3. the section of colloquial conversation in the reference Spanish corpus CREA (Corpus de Referencia del Español Actual), featuring about 12 million words (Romero-Fresco 2009: 52).
Results show that there is a close proximity in the frequency of the features investigated between Spanish original audiovisual dialogue and spontaneous spoken language. By contrast, the language of dubbing exhibits traits of unnaturalness and shifts towards the written and formal pole of variation, in part due to the adherence to internal dubbing norms as well as to the calquing from the English original texts. Interestingly, by comparing the occurrence of *de veras* in the Latin American and European components of CREA, Romero-Fresco argues that the unnatural frequency of the intensifier in dubbed Spanish is a vestige of *español neutro*, a standardized form of Spanish used when products were first dubbed in Latin America for all Spanish-speaking audiences.

Following a similar, predominantly target-oriented approach, Pavesi (2008, 2009a) examined the alignment between translated dialogues and spontaneous spoken language in her study of five syntactic and pragmatic phenomena in Italian dubbing. By comparing the frequency of features such as the conjunction and pronoun *che* ‘that’ and personal pronouns in a small corpus of dubbed Anglophone films and in the LIP (a reference corpus of spoken Italian), the author argues that naturalness is conveyed to viewers metonymically by giving priority to ‘privileged carriers of orality’ (Pavesi 2008: 79; 2009a: 209). Pavesi (2009b) widened this comparative approach in a follow-up study on the distribution and function of personal pronouns in the PCFD. Besides contrasting translated dialogue with spontaneous conversation and with original Italian films in the Forlixt1 corpus (Valentini 2008, Heiss and Soffritti 2008), the author compared the frequencies in the PCFD sub-corpus of original English dialogues with data extracted from a reference corpus of English conversation. These multiple comparisons revealed that the influence of English patterns on the translated dialogues is genre-specific and promotes the overrepresentation of second person pronouns in the language of dubbing.

As already highlighted by Romero-Fresco (2009, 2012) and Pavesi (2008, 2009a, 2009b), corpus analysis is indeed more informative when both translated and domestic screen dialogues are systematically compared to the language variety they are meant to approximate: spontaneous spoken language. The approach is thoroughly illustrated by Baños (2014a) in her target-oriented investigation of a broad selection of major spoken features belonging to the phonological, morphological, syntactic and lexical levels in a few episodes of the Spanish dubbed version of *Friends* and the domestically produced Spanish sitcom *Siete Vidas*. The results suggest that domestic products align more with colloquial conversation than do comparable dubbed products. This finding is shared with other corpus studies on dubbing including Rossi’s (1999) pioneering research on more than 100 spoken features in neo-realistic films in Italian and Matamala’s (2009) investigation of the interjections contained in sitcoms dubbed and originally shot in Catalan. If it is true that only corpus data can reveal systematic similarities and differences between original/domestic and translated audiovisual products, the results obtained in the above-mentioned investigations will need to be validated by means of bigger corpora that contain more varied selections of audiovisual products.

Corpus-based studies of subtitling have also addressed the typicalities of the specific translation modality, instantiated in this case by brevity and the representation of speech in writing (Tirkkonen-Condit and Mäkisalo 2007, Mattsson 2009). More specifically, according to Tirkkonen-Condit and Mäkisalo (2007), since the language of AVT is a type of translated language, the alignment between it and other translated languages is worth investigating to assess the impact of the specific medium and modality of transfer. In particular ‘[t]here are reasons to expect . . . that subtitle language might differ from other varieties of
translated language in its exploitation of such cohesive devices that contribute to brevity
and conciseness’ (Tirkkonen-Condit and Mäkisalo 2007: 222). The research starts from a
subtitle corpus totalling about 100 million words, the Finnish Broadcasting Corpus (FBC),
and the comparable Corpus of Original and Translated Finnish (CTF), developed at the
University of Savonlinna and counting up to 10 million words. From these sources three
sub-corpora were used in the study:

1. all the Finnish subtitles contained in the FBC with reference to the year 2004 (5.9 mil-
   lion words);
2. the Translated Finnish sub-corpus of the CTF (5.9 million words);
3. the Original Finnish sub-corpus of the CTF (3.8 million words) (Tirkkonen-Condit and

Clitic particles and cohesive devices were chosen to test the hypotheses about the specificity
of subtitle language. Among them, -kin is a brief and multifunctional particle conveying sev-
eral rhetorical relations of addition, elaboration and contrast, while -hAn is a multifunctional
discoursal device carrying an implication of shared information. Results showed that those
cohesive devices that are short, dialogic and unique to Finnish are more frequent in subtitled
Finnish than in translated Finnish and occasionally non-translated Finnish as well. In other
words, the brevity and colloquiality of the cohesive devices acquire functional priority in
the language of subtitling, thus overruling the Unique Item Hypothesis (Tirkkonen-Condit
2004), according to which target-language-specific items will be less frequent in translated
language owing to the lack of a translational stimulus in the source language. For example,
-kin is more frequent in subtitled Finnish than in translated Finnish, while -hAn occurs more
frequently in subtitled Finnish than in both translated and original Finnish. These results
corroborate the authors’ hypothesis that the language of subtitles is an autonomous variety
of the target language.

Finally, the language of AD is different from the language of other modalities of AVT
since it is not built to simulate spoken dialogue. The corpus-based study of AD has in fact
concentrated on the language of description and narration, given the specific functions
this intersemiotic translation performs within audiovisual texts. Plot-propelling elements
such as characters’ appearances and events must be represented linguistically to become
accessible to blind and visually impaired spectators. Through frequency counts, Salway
(2007) found that AD is characterized by nouns referring to characters, body parts and
concrete entities (e.g. ‘man’, ‘head’, ‘door’ and ‘room’), as well as by verbs of actions
conveying ‘manner’ (e.g. ‘hurries’, ‘leans’, ‘walks’, ‘smiles’, ‘watches’, ‘grabs’). The
higher than average frequency of these forms sheds light onto the focus of AD, whose
function is to provide concise ‘information about events in cause-effect relationships
occurring in space and time, and about the characters involved in the events and their
emotional states’ (Salway 2007: 154). The higher than expected frequency in the TIWO
corpus of the verbs ‘stops’, ‘starts’, ‘begins’ and ‘finishes’ additionally shows the rel-
evance of temporal transitions in this AVT modality, while the phrases ‘looks at’, ‘looks
up at’, ‘looks down at’ and ‘looks around’ introduce information about characters’ centre
of attention (Salway 2007: 160).

All these corpus studies empirically support the autonomy and register-specificity of the
languages of AVT in the main modalities in which they are realized and have been so far
investigated.
Translation tendencies: norms and universals in audiovisual translation

Corpus-based investigations of AVT have also been concerned with the search for translational norms and translation universals to suggest generalizations on the language of AVT and infer the processes underlying translation outcomes. Through quantifications allowed by corpus analysis, a few studies have explored the role of standardization, interference and simplification (Laviosa 2009), together with routinization vis-à-vis creativity (Pavesi 2016), in defining genre- or register-specificity.

Formentelli (2014) investigated whether lexical simplification impinges on dubbed language both affecting the transfer from source to target texts, and applying to the comparison between translated and non-translated dialogues in the same language (cf. Chesterman 2004). Simplification was defined along several parameters that allowed the researcher to explore the complexity and variability of vocabulary choices. Four randomized samples were extracted from the PCFD and were annotated so as to allow a variety of frequency measurements. Results suggest that lexical simplification does not occur during the process of translation from English into Italian. By contrast, lexical simplification emerges from the analysis of core vocabulary: that is, non-translated Italian dialogues contain numerous instances of non-standard and regional terms that do not occur in the Italian dubbed dialogues.

The role of interference, on the other hand, was explicitly tested in a study on primary interjections in the PCFD translational component supporting the Unique Item Hypothesis (Bruti and Pavesi 2008). Italian interjections that exhibit a degree of similarity with English ones (e.g. oh, ‘oh’/ ‘ah’) tend to be over-represented in dubbing, whereas those that are specific and restricted to Italian tend to be under-represented (e.g. multifunctional eh ‘hm’/‘oh’ etc.). Related to the law of interference (Toury 1995/2012) are also a few corpus-based studies that deal with translational routines, i.e. reiterated translation solutions generating recurrent strings in the target texts. Pavesi (2005, 2008) looked at such stock translations as the marked, non-standard but spoken constructions that repeatedly render English spoken constructions in dubbed Italian, whereas Freddi (2009, 2012) examined the extent to which routinization in the PCFD was carried over from the source into the target language when translating naturalistic versus plot-developing formulas. Confirmation of the productivity of translational routines in dubbing comes from Zanotti’s (2014) study of general extenders like ‘or something’ and ‘and stuff’ in a DIY corpus of American series dubbed into Italian.

A closer look at translation strategies and overall distributions

By means of bilingual concordances and corpus annotation, researchers have further shed light on translation strategies and shifts both in dubbing and subtitling (e.g. Tirkkonen-Condit and Mäkisalo 2007, Baños 2013, Matamala 2009, Pavesi 2013, Mouka et al. 2015). These studies have focused on typical features of orality such as discourse markers and vocatives (e.g. Rica Peromingo 2014, Forchini 2013) to identify translation regularities in translation operations, among which omission, equivalence and substitutions. A good example of corpus-based research on translation strategies is Formentelli and Monti’s (2014) investigation of the translation strategies used in transferring slanguage from English into Italian. The study confirms the trend towards standardization and neutralization of informal and taboo vocabulary in dubbing. Of the 1,219 instances of English slang words, dirty words and swear words extracted from the PCFD, almost half are rendered by functional equivalence.
in Italian, whereas the greatest share (54 per cent) are either downgraded or omitted. Further evidence of neutralization or standardization is provided by an investigation on the dubbing of intensifiers from English (e.g. ‘so’, ‘really’, ‘very’) into Spanish (Baños 2013). In this case, however, the search of the parallel corpus made up of 10 episodes of Friends showed that most of the standardized choices in dubbing mirrored the structure of the source text, presumably for reasons of synchronization, giving rise to more conventional syntactic combinations of intensifier + adjective, e.g. *muy importante*, ‘very important’. In the translations, however, there were also instances of more colloquial devices like morphological augmentation (e.g. *feísimo*, ‘very ugly’) and lexico-semantic resources, which selectively contributed to the naturalness of the translated audiovisual product. The tendency to naturalization in dubbing was in turn substantiated by the analysis of the operations carried out in the translation of English demonstratives into Italian (Pavesi 2013). In keeping with the optionality of the grammatical subject in Italian, English demonstratives in the PCFD were deleted most frequently when in subject position, with 42 per cent of subjects ‘this’ and ‘that’ disappearing from the dubbed Italian component. Substitutions (22 per cent of all translations) also contributed to bringing the target texts in line with target language preferences, while compensation strategies unveiled equivalent resources available in Italian speech. Cumulatively, these strategies were found to radically change the representation of deixis in the target dialogues. Whereas in English demonstratives signpost salient entities in the scene and in narration, in Italian access to the same entities must occur directly without the mediation of verbal language, as exemplified by ‘This is the only thing you got?’ translated with a subjectless clause in Italian: *È l’unica cosa che hai?* ‘Is the only thing that you have?’ (Pavesi 2013: 109).

While parallel corpora are clearly necessary to investigate translation strategies, they can also be fruitfully used to make comparisons between the overall distributions of features in source and target texts. A series of investigations on greetings, leave takings, introductions and good wishes (e.g. Bonsignori et al. 2012, Bonsignori and Bruti 2015) explored the role of these conversational routines in audiovisual products from a pragmatic, filmic and cross-cultural perspective. The comparisons between frequency lists (see above) brought to the fore the Italian formulas that are systematically favoured in dubbing and in subtitling to express the same functions as the English ritualistic speech acts. Focusing instead on dialogue-structuring syntactic phenomena, Ghia (2014) studied interrogatives contrastively in English and dubbed Italian dialogue. All direct questions in a sample of both the original and translated components of the PCFD were tagged for form and function, distinguishing among categories of *wh* - vs. *yes-no* interrogatives, and information-seeking vs. pragmatic questions typically used to challenge the interlocutor (e.g. ‘What are you doing?’). Like the original English dialogue, dubbing was found to be typified by a high frequency of interrogatives, in particular those that foster conflict and contribute to plot development. By contrast, on a syntactic level, questions in translation align with the target language and reproduce typical patterns of Italian conversation.

The studies reported above corroborate the usefulness of CBAVTS in identifying general trends in translated texts. Through such corpus findings, AVT is confirmed to result from several thrusts co-present during the transfer process. This means that the search for general tendencies and the study of translation strategies in AVT are inevitably interconnected with each other, whilst both impinge on the definition of naturalness. It also means that a major methodological difficulty in corpus-based research is isolating and separating the various components that shape AVT and its language.
Future trajectories

Corpus-based studies are a powerful approach to explore the dynamics of the process and the nature of the product of AVT both systematically and in-depth. There is, however, ample room for development in the field as less orderly work has been carried out in AVT in comparison to other areas of CBTS. Among the methodological weaknesses of current CBAVT research we find limited corpus size, partial comparability between sub-corpora and partial exploitation of the potentials of computer-assisted analysis. Researchers need to address these limitations to guarantee more reliable and representative datasets as well as more varied and conclusive research findings. The lack of bidirectionality in available parallel corpora represents an additional drawback when we attempt to establish cause-effect relationships in corpus analysis. Moreover, bilingual parallel corpora, which only have one source language (mostly English) and few target languages (mainly Italian and Spanish) limit the investigation of features that are specific to translated texts (Saldanha and O’Brien 2013: 68), thus hindering the search for translation universals in AVT. Hence, more source and target language pairs need to be involved and multilingual corpora should be compiled in order to gain data from a larger set of source language-target language combinations. The broader implementation of diachronic corpora would be an additional asset to research allowing analysts to probe challenging questions including how translational norms become established and how they evolve in time in different linguacultural settings and across spatial divides.

As for the phenomena investigated, there is still a lack of major quantitative investigations providing overall profiles of translated languages in all transfer modalities (cf. Pérez-González 2014: 117). Researchers thus need to move beyond limited descriptions of individual features to tackle the wider picture of AVT. The more systematic reliance on corpus-analytical resources and statistical measurements is also necessary to validate the significance of the findings, gain new ones and extend them to various aspects that have been so far under-investigated in CBAVT research. These include evaluation and ideology, which can be profitably explored by such means as semantic and pragmatic annotation, keyword lists, semantic prosody and sentiment analyses, as exemplified in Mouka et al. (2015). Not to be neglected are unsolved copyright issues as they jeopardize corpus creation itself and seriously compromise access to corpora, a limitation hampering the sharing of resources and replicability of results. At the same time, new technological applications should be envisaged including the larger use of research innovations that allow automated dialogue transcription, tagging and synchronization between text files, as well as between text files and audio or video files (Baños 2014b) in the transformation of monomodal corpora into multimodal corpora.

Besides representing instruments for descriptive and explanatory advancement in AVT studies, translation-driven corpora are liable to more extensive applications in various areas of the wider field, among which quality assessment and most importantly translator training, as shown in Sotelo Dios (2015). They can also be exploited in related research areas comprising Contrastive Linguistics, Intercultural Pragmatics and Conversation Analysis, together with Second Language Acquisition and Foreign Language Teaching (Rica Peromingo et al. 2014). These far-ranging applications additionally testify to the ample potentialities of CBAVTS and deserve to be explored in further enquiries in the field.
Summary

A corpus is a principled collection of texts stored electronically to be analyzed quantitatively and qualitatively by using computer-assisted techniques. In this chapter we have shown that CBAVTS belong to the wider approach of CBTS whose premises and methodological tools they share. Spoken and written corpora of AVT can be systematically searched to obtain generalizations about the make-up of translated texts and the process of translation by investigating frequencies, recurrent word combinations and collocations, along with cross-linguistic and cross-cultural correspondences. This approach has allowed researchers to address several important issues including the naturalness of translated screen dialogue, translation strategies, general tendencies and variation in translational behaviour, thus proving to be a powerful epistemological resource in AVT worthy of continual development and refinement.

Further reading


Related topics

3 Subtitling on the cusp of its futures
4 Investigating dubbing: learning from the past, looking to the future
8 Audio description: evolving recommendations for usable, effective and enjoyable practices
13 Spoken discourse and conversational interaction in audiovisual translation
18 Sociolinguistics and linguistic variation in audiovisual translation
21 Multimodal corpora in audiovisual translation studies
22 Eye tracking in audiovisual translation research

References


PCFD (n.d.) ‘Pavia Corpus of Film Dialogue’. Available online: http://studumanistici.unipv.it/?pagina=p&titolo=pcfd [last access 20 December 2017].


**Sitography**

AntConc 3.2.1w: www.laurenceanthony.net/software/antconc/ [last access 20 December 2017].

Wordsmith Tools: www.lexically.net/wordsmith/ [last access 20 December 2017].