Multimodal corpora in audiovisual translation studies

Publication details
Marcello Soffritti
Published online on: 06 Sep 2018
21

Multimodal corpora in audiovisual translation studies

Marcello Soffritti

Introduction

Since the 1990s, corpus-based research has been the drive behind key methodological advances—not least the development of techniques to quantitatively interrogate large computer-held collections of texts—in the neighbouring fields of linguistics, lexicography and translation studies, to name but a few examples. Corpora have become powerful and reliable tools to compile representative samples of authentic texts pertaining to various language varieties and genres, test hypotheses regarding the frequency and regularity of certain linguistic patterns that cannot be verified by the researcher’s intuition, and support the generalization of linguistic findings. Effectively, corpora expose the limitations of linguistic research based on made-up examples and the analyst’s subjective judgement.

This chapter examines how multimodal corpora (MMC) can be exploited for the purposes of AVT research. It discusses the architecture of these resources, consisting of audiovisual texts and their translations, and explores various aspects pertaining to the design, compilation, processing and maintenance of these corpora. After examining some of the difficulties that are endemic to MMC, the chapter ends by suggesting different avenues for the development of such computer-held resources in terms of their contribution to AVT research.

Types of corpora in translation research and practice

There are some obvious differences in the way translation researchers and practitioners generally use corpora. Translation practitioners are result-oriented and concerned with working out appropriate translation solutions for real life problems, often bound by a range of technical and organizational constraints. For them corpora are chiefly tools to carry out professional assignments under defined conditions, often in combination with a number of other tools, to enhance their productivity as translators. Dedicated corpora for the translation industry often take the form of translation memories (TMs), i.e. databases typically consisting of aligned sets of previously translated (source and target) text segments that translators can retrieve and edit as they translate new texts bearing a varying degree of resemblance with those stored in the database. TMs can be considered as parallel corpora.
whose representativeness is limited to mostly technical and professional domains, and whose relevance to the translator’s work is maintained through continuous refinements and enlargements of the database every time new validated translations are added to the memory. In spite of sharing some fundamental characteristics with parallel corpora, TMs are not particularly helpful for research in translation, because they do not generally enable researchers to identify and explore strategic and methodological issues of scholarly and pedagogical relevance. Instead, translation scholars have benefited from the compilation of large annotated corpora, whether monolingual or bilingual, general or specialized.

Recent trends in translation theory and multimodality (Pérez-González 2014) call for a more systematic study of the contribution that non-verbal semiotics make to the overall meaning of texts combining written/spoken language, sounds and images. Just as linguistics and translation theory have capitalized on the growing availability of corpora, research in multimodality has become increasingly dependent on corpus-based evidence to verify and refine its own basic assumptions and statements (Bateman et al. 2002). Against this backdrop, this chapter explores the current and potential capacity of MMC to yield new insights into AVT. In contrast to their unimodal counterparts, multimodal corpora contain texts made up of both verbal and non-verbal meaning-making modes, and hence encompass the same kind of texts that lie at the heart of the AVT process. Ultimately, MMC are required because setting up ‘a database or a corpus of film dialogues and their subtitles, with no pictures, and still pretend to study screen translation’ would be pointless (Gambier 2006: 7).

The emergence of MMC can be traced back to the turn of the century, although calls to include audiovisual components in electronic corpora date back even earlier. A list of multimodal databases in López-Cózar Delgado and Araki (2005: 229–231), very likely out-of-date at the time of writing, quotes about 20 instances, not all of which are available on the Internet. Their number has probably increased since it was first published, although it will certainly remain negligible when compared with the number of unimodal corpora available to language scholars.

As multimodality establishes itself as a new qualitative dimension in the compilation of corpora, by incorporating non-written and non-verbal material, the written representation of language loses its primacy. The increasing number of communication channels represented in such corpora and technological platforms required to access them also add to the variety of multimodal data to be collected and stored (Knight 2011: 405). In the case of MMC consisting of audiovisual texts and their translations, corpus design is particularly challenging, because they must include multimodal content in both the source and target languages. The advent of digitization has, however, expedited the compilation process: AVT corpora can now be built more easily due to the greater availability of digital materials, the widespread availability of tools to compress digital data, and the growing ubiquity of suitable storage devices and online platforms for the exchange and distribution of audiovisual material. Another important factor to bear in mind in the era of social media is the willingness of users and communities to contribute materials in numerous ways—even though the quality and relevance of crowd-sourced materials are often questionable, and hence have to be carefully checked before they are included in the corpus.

**Mining the Internet for multimodal corpora**

Quite a different situation emerges if one looks at trends in mainstream web-based communication. Countless web pages are multimodal to varying degrees, in that they include at least visual elements conveying meaning, such as static pictures, graphics, symbols, icons,
menu bars in different colours. In addition, various sounds (music, sound effects, and even human or animal voices) are often deployed in combination with written or visual elements. Acoustic information is sometimes available in separate files for retrieval. The same applies to motion videos: recordings of people, objects, events and locations can be integrated into websites, sometimes as constitutive components of the wider text, sometimes as autonomous items for purchase, practical use or simply as complementary information—whether it can be accessed directly or via a dedicated link.

The translation of webpages, as defined above, raises important questions that do not apply in other text types. Are translations of webpages to be considered as a form of AVT? And to what extent are they amenable to being included into a corpus of sorts? A good reason for including them in this discussion lies in the fact that, in many websites, versions of certain pages are readily available in other languages. Indeed, in some cases these translations exhibit changes in content vis-à-vis the original, as they are tailored to the needs of different audiences in the target locales. As a general principle, MMC informing language and translation research should consist of couplets of existing source products and their existing validated translations. Multilingual versions of webpages are therefore potentially relevant and could be considered for mining and inclusion in MMC. Wikipedia articles in different languages, including both verbal and non-verbal meaning-making resources, are a case in point.

On a different quality level, Internet browsers now feature a (machine) translation function that can be directly applied to most standard webpages. Translated web pages are thus the output of a corpus-based process, where the translation solutions are extracted from a corpus ‘hidden in the cloud’. Specifically, they are generated by machine translation engines that process huge collections of parallel text segments, feeding on millions of webpages and owned by leading companies (Google Translate, Google Translator Toolkit, and some others in the recent past), in what represents the most prominent example of corpus-based translation supported by large-scale harvesting of parallel texts. Neural computation, statistical power and more or less conscious crowd involvement in an ever-ongoing process of engine refinement are the basis for this largely non-linguistic corpus building methodology. To a considerable extent, web-based machine translation is capitalizing on the growing reliability and ubiquity of special tools developed under the umbrella of existing language corpora (e.g. lemmatizers, parsers, part-of-speech taggers), and the qualitative output of automatic translation engines is increasingly appreciated by both average and specialized end-users. Crucially, the growing involvement of the crowd in the generation of data is likely to enhance and refine the amount of validated translations within such parallel ‘corpora’.

In order to decide whether, and to what extent, machine translations of webpages can be regarded as MMC, one must take a closer look at the objects and processes at stake. MT-processed texts feature within webpages that also contain non-verbal elements, i.e. meaningful visual and/or acoustic content. However, only the written text within those multimodal ensembles is processed by MT engines, since the underlying corpus consists exclusively of segments of written language. Similar or even stronger limitations would appear to apply in most tools used to localize video games. It goes without saying that the new generation of such tools is unlikely to remain bound by such technological constraints.

More complex operations of the same kind take place when we look at another mainstream translation process involving multimodal material, i.e. the automatic translation of subtitles available for a growing number of YouTube clips. Subtitles available in this context are ever more frequently being produced automatically, i.e. with the help of speech
recognition applications that manage to covertly perform written transcriptions of spoken words in real time. The output of speech recognition software is then promptly processed by MT applications that generate target language subtitled versions. Speech recognition and MT, both of which are based on a combination of statistically arranged data and increasingly sophisticated conversion rules, work in hand-in-hand.

Both MT engines and speech recognition software intersect with the definition of MMC presented in this chapter. Although the role played by textual corpora is more evident in the case of MT, speech recognition is probably closer to multimodal analysis, in that it applies segmenting and combining rules that are based on large inventories of spoken documents. These have been filtered so that they are representative of as many phonetic and acoustic elements as possible, which are in turn interwoven with rules of orthographic transcription. The deployment of two consecutive automatic processes of content conversion obviously entails a higher risk of error chains caused by the incorrect decoding of inaccurate or unclear pronunciation of the source words, along with a number of other imponderable shortcomings that can occur in later stages of the process. User satisfaction is still low and probably will remain so as long as the accuracy of automatically generated target subtitles remains below 90 per cent. However, insofar as this service is free and it is likely to improve, at least for some of the most widely used languages, this form of translation is not likely to disappear any time soon.

Some scholars will have reservations about the fact that machine translated webpages rely on procedures applied to simple text chunks and, as a result, the relationship between these mechanically produced target versions and the audiovisual material embedded in the text may be seriously disrupted. Still, the translation is not necessarily completed after a single iteration of this process. Rather, it can undergo one or more subsequent stages of validation; crucially, motivated end users can optimize these results in terms of the overall multimodality of the end product—if it is still required.

In sum, large-scale processes of translation driven by big data tend to rely on unimodal tools resembling ever growing parallel corpora, even for multimodal webpages. This simplified but highly effective approach to translation cannot be ignored nor hastily dismissed by cooperatively minded AVT practitioners. There are strong reasons to share methods or experiences: multimodal products, corpora and AVT are highly interdependent and more pervasive and, given their flexible boundaries and degree of overlap, they cannot be considered as separate fields.

Multimodal corpora

Foster (2007: 1) defines MMC as ‘a recorded and annotated collection of communication modalities such as speech, gaze, hand gesture, body language, generally based on recorded human behaviour’. This definition obviously draws on the classical definition of multimodal texts as those ‘whose meanings are realized through more than one semiotic code’ (Kress and van Leeuwen 1996: 183, but see also Allwood 2008: 207 ff.). On the other hand, the term ‘multimedia corpus’ is used to designate ‘a systematic collection of language resources involving data in more than one medium’ that normally ‘consists of a set of digital audio and/or video data and a corresponding set of textual data (the transcriptions and/or annotations of the audio or video data)’ (Schmidt et al. 2014: 1). Given their shared conceptual core, this chapter will treat both terms as synonyms and will henceforth refer to both as MMC, thus reflecting the growing tendency to adopt an enhanced notion of multimodality (Bednarek 2015: 65).
The scholars cited in the preceding paragraph conceptualize MMC as monolingual collections enabling research on the complexity of meaning-making activities—mostly in dialogic settings, but also in human–machine interactions—through a growing range of advanced multimodal user interfaces (see Abuczki and Esfandiari 2013 or, for a more comprehensive charting of this area of research, the range of papers published in the *Journal on Multimodal User Interfaces*). Most of these interfaces contain and display annotated textual records of spoken communication consisting of sound recordings, video-streaming (as these interactional encounters are made with purposefully arranged cameras and microphones) and (written) transcriptions of the verbal content. Annotation is carried out in each of the three channels that run simultaneously when the user examines an extracted sample. It addresses not only classical features of syntax, vocabulary, morphology and style (on the written channel), but also intonation, pronunciation, pitch, pauses, and para-verbal signals (on the acoustic channel), as well as facial expressions, gaze, body movements and camera shots (on the visual channel). All annotated units are equally relevant, and they should all be ideally available when the corpus is being interrogated, although text search probably remains the quickest and most efficient way of locating contents in a corpus.

Hits generated in response to specific search queries are generally presented as multimodal concordances, which deliver a comprehensive overview of which meaning-making elements concur in a single segment of the dialogue (see, for instance, Baldry and Thibault 2008). Another key source of information lies in the statistical values generated by relevant aspects of the database architecture. These include quantitative information about annotated elements and the combinations, patterns and profiles that they are part of; comparisons between MMC data and data from general purpose corpora; and many more findings that can significantly enhance our knowledge of how communication works in different settings and how it can be best analyzed. Even research on the interaction between humans and technical devices in controlled situations can benefit from the interrogation of MMC: gestures, facial expressions and the emotional state of humans can be modelled and ultimately simulated so as to build artificial agents with the capacity to ‘understand’ and speak sensibly to human partners, or to assist impaired persons.

Being mostly monolingual, these MMC are only tangentially relevant to AVT because they do not feature comparable data in two or more languages—although Navarretta et al. (2011) present a rare instance of comparable MMC for some Nordic languages. To facilitate AVT research, corpora must contain source and target language multimodal texts. In other words, AVT-relevant MMC should include most features of monolingual MMC plus bilingual or multilingual aligned language data, together with a facility to run various query modalities in the user interface. In corpus linguistics terminology, these MMC of audiovisual texts and their translations should ideally be parallel. By embracing original audiovisual products together with their translated counterparts in one or more additional languages, these MMC respond to the definition of parallel corpora as collections in which ‘two or more components are aligned, that is, are subdivided into compositional and sequential units (of differing extent and nature) which are linked and can thus be retrieved as pairs (or triplets, etc.)’ (Fantinuoli and Zanettin 2015: 4). AVT translation is also heavily dependent on comparable corpora—typically a pair of corpora in two different languages which come from the same domain but are not translations of each other. However, given the notorious difficulties entailed in determining the optimal composition of comparable corpora (Fantinuoli and Zanettin 2015: 4), this chapter focuses exclusively on parallel corpora.

Contrary to what happens in unimodal and monolingual corpora, in MMC the target element of an original-translated couplet is processed not only in terms of its verbal components,
but also in terms of its acoustic and/or visual ones, occasionally shortening or stretching the length of translated sequential units in comparison to their sources. On the other hand, like unimodal and monolingual corpora, multilingual MMC can differ significantly according to their intended use. Corpora can be used as a resource to inform scholarly research (e.g. research in AVT methodology, applied AVT, certain aspects of intercultural communication, semiotics of translation, etc.) or as a semi-practical tool in professional AVT training, e.g. as a repository of possible solutions to translation problems. The latter use is relevant to dubbing/subtitling in contexts where teamwork, standardization and reusability of the validated solutions are required.

**Multimodal corpora for audiovisual translation studies**

Unlike webpages, audiovisual products are generally identifiable as such in that they can totally or largely dispense with written components. They are made to be viewed and listened to, rather than read. Audiovisual texts also differ from webpages in that they are not necessarily located on the Internet, as they can be generated, distributed and used offline. Prototypical audiovisual products are also the outcome of traditional and autonomous industries and activities such as cinema, TV programmes, pedagogical publishing, theatre, and advertising. Newer and widespread audiovisual textualities include videogames, tutorials, documentation for technical and professional purposes, product information, and more. And since this list does not yet account for genres or distribution modalities, additional distinctions between classes of products may become necessary according to the translation modality (e.g. written, intersemiotic, dubbing, subtitling, voice over) used in each case. Accordingly, MMC for translation generally need to take into consideration a large number of different factors.

**Multimodal corpora for translation**

Unlike their general counterparts, parallel corpora used by translation scholars are not fully representative of general categories, such as genre, subject, or time period. Building a parallel corpus to the British National Corpus (BNC), for example, would involve finding validated translations for every text held in it. Even if this were possible, that parallel corpus of translated texts would not meet the requirements to be regarded as representative of the target language in question. Parallel corpora are at best coherent and homogeneous subsets of texts chosen to serve specific scientific, pedagogical or professional purposes, and these intrinsic limitations are particularly acute in the case of MMC consisting of audiovisual texts and their translations.

In this light, the appropriateness and usability of parallel corpora will depend on whether they are meant to support research on speech functions and translation quality, to give but two examples, or alternatively they are expected to provide operational support to translation process, translator training or even translation quality assessment. These factors are decisive in optimizing the design, expansion and modes of exploitation of parallel corpora. As a general requirement, however, parallel corpora conceived to assist translation scholars with their research should hold two language versions of the same text along with some sort of alignment function. Alignment, however, might not always be an option. Although visual data tend to remain the same in most translations of texts combining language and non-verbal meaning-making resources, some webpages may undergo processes of adaptation or localization to ensure that the visual and/or acoustic content is adapted to the needs or
expectations of particular target groups. Likewise, movies, TV programmes, ads and other forms of audiovisual content may require structural changes during the translation process (Gambier 2006). Should a deep reshaping be required on all channels, these audiovisual texts may become impervious to alignment as found in unimodal parallel corpora.

MMC can hold certain types of audiovisual texts that are (or are to be) translated. Along with films, several TV genres (e.g. documentary films, cartoons, reports, interviews etc.), theatre performances (including opera and musicals), videogames, interactive virtual assistants, pedagogical materials and audio books are likely to be of commercial or political interest, and hence likely to be translated. In the case of MMC consisting of audiovisual texts and their translations, current corpus standards dictate that non-digital objects and data be stored in a digital format. Additionally, a written transcription should be provided and added to the audiovisual material, to make verbal elements of speech readily retrievable.

Aspects of multimodal corpus design, building, and exploitation

To develop an MMC that supports AVT research, general management protocols (Hunston 2009) and guidelines for the processing of specific components or features of the corpus are required. A detailed and critical description of the general tasks involved in corpus building is available in Lüdeling and Kytö (2008–2009). What follows is a selection of the main points to be considered specifically in the case of MMC made up of audiovisual texts and their translations. These points will be presented without delving into the technicalities of database architecture, application design, operating systems or browser requirements. Instead, the main focus of this overview concerns what is typically needed to build parallel MMC of original and translated audiovisual texts to assist with translation analysis and, to some extent, with quality assessment. This reflection on the specific complexity of MMC of audiovisual texts and their translations will foreground a range of fundamental methodological issues and enhance awareness of the peculiarities, opportunities and limitations arising in related empirical research.

Among the specific problems concerning parallel corpora for AVT, representativeness is the first to be tackled in the planning stage. The representativeness of MMC for AVT must be established on the basis of a larger array of parameters than the one informing the compilation of unimodal or monolingual corpora. Large institutional corpora of written language are considered as representative when their size exceeds 100 million words (the BNC consists of 100 million words, the DWDS corpora holds more than 2 billion words, while DeReKo exceeds 25 billion words). Institutional corpora of (transcribed) spoken language, on the other hand, are much smaller and none can claim to be fully representative of a given language. This inadequacy of spoken language corpora is obviously due to the amount of time and resources required for transcription and annotation, but also to various difficulties pertaining to the gathering of sufficient speech recordings. Applying the same representativeness criteria to MMC is not feasible: the structural complexity and manifold alignment problems associated with MMC would require an unmanageable volume of material and amount of data processing. Besides, parameters of representativeness should first be set for the various semiotic resources represented in audiovisual texts, including visual, acoustic, kinetic and generally non-verbal information, as described by multimodality theoreticians. But even assuming that the multimodal coverage of an entire given language could be plausibly planned and then seriously attempted, building a corpus meeting these specifications may still not
be viable, given the need for huge storage space and potentially uneven quality of the recordings included in the collection. Finally, an MMC supporting research in audiovisual texts and their translations is only as good as the amount of parallel couplets it contains. Insofar as such couplets are rare and often difficult to gather, the goal of achieving representativeness moves definitely out of reach.

MMC for AVT should therefore aim not at general, but simply at achievable representativeness, and consequently restrict the scope of the projected collection to subsets of entities selected by kind of object, genre, time span, cultural impact, seminal potential, etc. Selecting this audiovisual material to be included in a corpus also implies opting for one or more types of multimedia products together with their translated version(s); establishing the number of source or target languages; and choosing one or more types of translation, i.e. subtitling, dubbing, voice-over, etc.

As most projects involving the development of MMC for AVT aim to rigorously document multimodal communication in translation, they should be fed exclusively by multimodal texts and their existing translations, rather than purposefully commissioned ones. Even if such translations are available on the Internet, they often have commercial relevance and are subject to copyright restrictions. The feasibility of including them in a corpus should therefore be carefully examined, and careful consideration should also be given to any other relevant regulations concerning access to and use of the corpus by third parties.

Other tasks concerning data acquisition and maintenance (purchase, web crawling, manual or automatic transcription, spectrographic analysis, exchange of already processed documents, testing and validation rounds, etc.) would deserve more space than this chapter allows. Instead, the remainder of this section focuses on the pre-processing, segmentation, annotation and alignment of data, since the way in which these aspects are managed in the context of MMC differs significantly from unimodal and/or monolingual corpora.

Complex activities such as the acquisition, storage and processing of data generally require special interfaces, particularly when it comes to handling different electronic formats, segmentation, time code allocation, alignment and file annotation. These steps, important as they are, remain hidden to end users and can be performed with one or more dedicated tools such as ELAN, PRAAT, ANVIL, EXMARaLDA, XTrans, etc.—see Schmidt et al. 2014. Choosing and adapting such data entry interfaces to handle audiovisual texts requires very accurate planning. In most cases, these tools not only enable the pre-processing of the multimodal files to be entered in the corpus, but they directly facilitate matching query functions once the corpus has been established (see for example Knight et al. 2008, Kipp 2014). Regardless of the applications chosen, staff responsible for data entry must be already skilled and/or especially trained not only in the use of these tools, but also in the handling of tags and labels.

The transcription of spoken texts and the internal segmentation of multimodal documents are additional tasks to be performed during the pre-processing of MMC. At a macro-level, the segmentation of texts included in an MMC for AVT obviously concerns generally video files and text files. Inside video files, sound and spoken words may be treated as separate data, extracted and stored in audio files, and represented as spectrograms or other speech waveforms. Text files are obtained through manual transcription, the purchase of original scripts, or with the help of speech recognition software applications, and they can be structured either according to official transcription guidelines for spoken language, or simply by conforming to standard orthography (Bonsignori 2009). Segmentation applies insofar as the components of the source subcorpus must be matched with their target counterparts in the translated versions.
At a lower level, documents contained in a corpus are generally composed of smaller sections and parts, which must be used as reference points when it comes to locating significant elements. Many written documents are already segmentable on the basis of their own internal structure (e.g. summary, introduction, chapters, etc.), and additional segmentation can be performed automatically at paragraph, sentence and even phrase level. This is not possible if the document is a video file: segmentation criteria must be then set by the project manager, e.g. from one camera shot to the next, from one setting to the next, etc.—for an instance of segmentation in a small corpus of subtitled films, see Sotelo Dios 2011: 253. In sound track files, segments can extend from a longer pause in conversation to the next, but other criteria can also be adopted. At all levels, time codes play a decisive role, because after corpus interrogation users will need their ‘hits’ to be modally complete, i.e. they should bundle (in temporal alignment) all visual, acoustic and written information pertaining to retrieved segments of the multimodal content. Unlike written corpora, in which a query renders its hits inside chunks of text of a pre-defined size, an MMC for AVT is bound to render chunks taken from video and/or audio tracks as well. How to optimize segmentation according to such constraints and requirements is still an open question for MMC designers.

The next distinctive feature of parallel MMC to take care of is alignment: between linguistic data, between linguistic and acoustic/visual data, between linguistic data and spectrograms, between scenes and subtitles, etc. On the temporal axis, alignment normally means synchronizing separate sound and video tracks with the help of time codes, attaching in a separate frame the corresponding lines of written text. Alignment between source and target elements requires both sets to have been previously chunked in the same way, so that source chunks are unequivocally linked to target chunks. Structural analogies to commercial TMs are clear, but this more complex kind of alignment needs dedicated applications. These must ensure that source and target elements can be simultaneously viewed by the user. Of course, peculiar alignment problems arise when target products have been changed with an increase or loss of footage in audiovisual components (‘parallelism holes’). Such crucial spots must be handled separately and properly marked.

Annotation is the basis for effective and satisfactory data retrieval, since only annotated features of corpus materials can be specified and extracted in queries. And, of course, only on such a basis is it possible to identify statistical and distributional patterns of retrievable entities efficiently and reliably. There are several layers of annotation, depending on the analysis methodology, nature of documents, scientific or professional approach, previous conventions, existing norms, manual vs. automatic annotation, critical quantities of annotations, sustainable variety of annotation, etc.).

Annotation involves marking up segments of the corpus according to an organized set of tags and labels, each of which delimits the places where a defined feature is found. As a rule, annotation must identify whole documents belonging to the corpus, stating relevant information on, for example, textual genre, author(s), provenance, time of creation, keywords, person or entities involved and much more. This requirement must be met to ensure that references and macro-categories are correctly attributable to lists and samples extracted from the corpus. Many of the labels relating to structure and content of documents can be assigned automatically or semi-automatically, but at lower levels, manual processing by skilled operators is quite often needed. This is one of the most crucial points in the pre-processing of MMC for AVT. Their low-level annotation must deal with more distinctive and fine-grained linguistic, stylistic, thematic, visual or acoustic features than is the case in unimodal corpora. These features are located in the objects at different distances from one another and on different semiotic dimensions. This makes annotation in MMC a crucial task
Multimodal corpora and AVT studies

to organize and requires a specific methodology, whose standards are not yet fully developed (for an example of tag used in film audio description, see Jiménez Hurtado and Soler Gallego 2013: 583).

Most importantly, annotation in MMC for AVT may have to be performed on a parallel basis as well, i.e. managing diverging sets of features for source and target elements. Actually, separate linguistic annotation of source and target elements should be performed for written and spoken components of the corpus, since different texts in different languages are involved. This means that, even using a common tag set, a source text chunk would be linguistically and acoustically tagged in a certain way, and its matching target chunk in quite a different way. Furthermore, couplets of matching elements belonging to source and target audiovisual texts may also be visually different, depending on how radically the visual information has been handled in the translation process. The more the visual elements are modified in the target version, the bigger the divergences will be in the annotation of the visual channel. Modifications of the video track are normally kept to a minimum in the case of expensive film or videogame productions, and are likely to consist mostly of the addition or subtraction of available footage. But there are other cases in which they occur more frequently, for instance in multilingual versions of news or documentaries broadcast by internationally acting corporations (e.g. BBC, Euronews, DW, Al Jazeera and the like). When such differences affect both visual and spoken information above a given threshold, there may be doubts as to whether it is still a parallel corpus.

Tags and labels must be planned in advance not only according to norms and conventions. They must also conform to general standards of solidity and consistency, because the

\begin{table}
\centering
\caption{Elements to be labelled in monolingual, dialogical multimodal audiovisual material.}
\begin{tabular}{ll}
\hline
1 & \textit{ACOUSTIC CHANNEL} \\
\hline
1.1 & Spoken language \\
1.1.1 & Para-verbal signs (not what is said, but how it is said) \\
1.1.2 & Voice qualities, cadence, inflection, or rate of speech. \\
1.1.3 & Distinctive sociolinguistic features and their connotations \\
1.2 & Non-verbal acoustic signs \\
1.2.1 & Music \\
1.2.2 & Special effects \\
1.2.3 & Sound arrangements \\
\hline
2 & \textit{VISUAL CHANNEL} \\
\hline
2.1 & Iconographic code \\
2.1.1 & Symbols and icons \\
2.2 & Photographic code \\
2.2.1 & Colour \\
2.2.2 & Camera angles and variations \\
2.2.3 & Focal length of the lenses \\
2.2.4 & Directorial editing choices \\
2.2.5 & Luminance and contrast patterns \\
2.3 & Mobility code \\
2.3.1 & Proxemic and kinesic issues \\
\hline
\end{tabular}
\end{table}
processing of corpus data takes place through one or more underlying databases, whose internal settings are pre-established and not changeable. With cautious database planning, adding labels at the end of a label tree may be allowed, if they should unexpectedly emerge after the start of the annotation activities. In any case, hierarchy, names and number of labels are governed by strict allocation rules (see e.g. several instances of annotation graphs in Schmidt et al. 2014).

Annotation of audiovisual texts in an MMC is in many regards tentative and experimental. One way to plan it could involve merging annotation schemes normally adopted for texts, audio and video documents, including suggestions from recent contributions by multimodality scholars (see Abuczki and Esfandiari 2013 for multimodal dialogue analysis; Maszerowska 2012 and Pérez-González 2014 in translation studies). Table 21.1 features a list of aspects to be labelled, as required for the analysis of ‘simple’ monolingual, dialogical multimodal material.

This already complex inventory is undoubtedly insufficient for full-fledged annotation. Other sets of labels and tags are equally important for the retrieval and analysis of the verbal components (lexemes, phrases, sentences, paragraphs, but also formulas, phrasemes, word collocations, n-grams, functional units in written and spoken communication, speech acts, puns, idiolects, dialects, regional varieties, fragments in foreign languages, and much more), and can be taken from classical corpora of written and spoken language.

Still more labels and tags are necessary for the management of relevant information concerning situations, contexts, temporal and spatial embedding, locations, cultural entities, or particular objects. It is extremely difficult to establish which and how many visual entities are actually relevant for the analysis of translation processes and the evaluation of translation quality, particularly when a corpus includes classes of visual products as news, fiction, documentaries, etc. Moreover, additional layers of tags may be required by particular translation processes of relevance to the corpus. When, for instance, dubbing or subtitling are involved, dedicated corpora may need mark-up of segments of video-shots in which lip movements are visible, or moments in which more than one person is speaking in the same subtitle frame.

Overall, the total number of tags to be taken into consideration easily goes into the hundreds (for descriptions of very rich tag sets in film corpora see Valentini 2009, Jiménez Hurtado and Soler Gallego 2013). Only a limited number can be handled automatically or semi-automatically, and the vast majority must inevitably be inserted by trained staff according to guidelines issued by the project managers. In projects involving more than one tagging person, intersubjective consistency in tagging non-linguistic entities is often difficult to achieve, given that definitions of numerous relevant phenomena are fuzzy, controversial or incomplete. Consistency among different projects would be even more difficult to guarantee. Tagging films or documentaries, for instance, implies taking note of countless entities, situations and events, distributed across different historical periods, geographical areas and cultural frames. So far, there has been no interdisciplinary agreement on how to code classes of visual objects (other than human figures) that may feature in audiovisual products to be translated. Even suggestions concerning the relevance of visual information in films are still to be transferred into precise tagging schemes. The same applies to classes of situations, interactional encounters between people as well as between persons and other entities, etc. All these elements are potentially relevant, and therefore require tagging, when it comes to checking regularities, constraints and quality in translation decisions and translated products (for examples and discussion of the methodological implications of complex AVT corpora interrogation enabled by rich tagging, see Valentini 2009).
Peculiarities of interrogation interfaces and result visualization will finally be considered briefly. Interrogation procedures generally empower users to specify as many combinations as possible among manually annotated categories together with automatically definable parameters. Since in an MMC there are numerous, heterogeneous annotated entities, particular care must be taken to allow flexible and transparent specifications of data to be retrieved, regardless of how complex a query can become. Extracting information from a richly annotated MMC is a crucial task in different regards. First, for every channel there are specific state-of-the-art procedures for displaying query results. New applications developed for large linguistic corpora, for instance, present a significant volume of important linguistic information distributed in many different windows, each of which may contain a table, a diagram, a graph or a word cloud. As seen in current interface models for unimodal monolingual corpora, interrogation can be often refined, narrowed, enlarged and re-structured in subsequent steps. Most importantly, relevant correlations between newly extracted data can be assumed, tested and established at different levels, including statistical distribution, syntactic and semantic patterns, and every kind of collocation or n-gram. These otherwise inaccessible, invaluable findings are frequently made possible by the availability of several windows at the same time. Second, this already wide range of linguistic data in different windows should be managed in light of additional requirements posed by a parallel, i.e. at least bilingual architecture, in which query results should be ideally displayed in parallel windows. Given the physical limitations of computer screens, such simultaneous visualization becomes a very difficult task, if one chooses to accommodate as many interrogation modalities as in state-of-the-art applications (AntConc, ConcGram, Sketch Engine) or reference monolingual corpora (BNC, DWDS and others). Again, such decisions shall be made according to general priorities concerning the intended use of a corpus, so that certain windows displaying translated chunks of verbal and/or visual material may be privileged for simultaneous visualization, and others displaying further information would be presented only on demand or after scrolling down. Third, a query on an MMC shall additionally refer to categories of non-linguistic data, as established in the annotation system. Depending on how many visual and acoustic labels are available, additional display options must be planned and accommodated in dedicated windows. Current multimodal concordances are a first step in this direction, but appropriate solutions for visualizing all these elements are still lacking.

Future trajectory and new debates

As their very low number clearly indicates, building usable MMC for solid empirical research in AVT is undoubtedly more complex and challenging than building unimodal corpora. Many methodological issues still require additional reflection, and acceptable solutions will probably only arise after additional trial-and-error rounds of pilot projects. However, even small and only partially annotated MMC for AVT have proved to be useful as an empirical basis for very many contributions, especially by Spanish and Italian scholars. They can be expected to become a starting point in the future for new and original investigations, beyond the traditional notion of ‘screen translation’. They could also be usefully exploited to calibrate to some extent new experimental procedures and tools of automatic annotation, along a path of mutual improvements in an interacting chain between intensive manual and extensive algorithm-driven operations.

Undoubtedly, methodological uncertainty and dependence on time-consuming manual operations put strong constraints on short-term enhancements. Most AVT researchers are well aware of these limitations and are likely to stick to limited multimodal translation
corpora for clear-cut purposes (including research on translation quality in specialized domains, advanced training in academic courses, or vocational training for companies and institutions).

Small existing MMC could in turn be assembled in bigger ones, but this option still faces unsolved methodological, technical and financial problems. To name just one, there are few (in some cases, not even one) parallel MMC available for several combinations of widespread languages, so no developments are likely to take place in the near future. Materials for comparable MMC would be easier to find, e.g. resorting to some multilingual TV-channels (as reported in Afli et al. 2014), but their usability would be limited to preliminary translation (or adaptation) steps.

Consolidating small MMC into bigger entities also raises issues of consistency. To achieve consistency and uniformity within a single corpus project, a very high price must be paid in terms of working hours, costs and overall control procedures. Consistency among different projects may even prove to be impossible to achieve, because the annotating effort required by academic investigation lies far beyond both the commercial and cooperative assets currently available. To deal with these limitations, it is probably inevitable to reduce the number of tagging operations in which interpersonal subjectivity is most critical, and try to rely as much as possible on automatic procedures. Also in this regard, large and consistent MMC should be developed through academic research in collaboration with big data owners, taking advantage of jointly developed automatic annotation. For now, however, there is a significant clash between what would be desirable to conduct rigorous multimodal analyses and what can realistically be performed manually and/or automatically. Considering the wealth of insights and serendipity that the advent of large corpora unleashed, the consolidation of small MMC into bigger ones would be highly desirable and allow for solid generalization of findings within several fields of descriptive linguistics. A new generation of large(r) MMC, apt to support AVT translation, may arise under certain conditions from vast collections of already available multimodal objects, harvested with advanced crawling techniques that are able to detect appropriate or necessary AV products in web-based repositories. These collections could in the future be made searchable and become at least partially usable as components of MMC for AVT as soon as automated speech recognition and automatic translation (for suitable audiovisual material) are so reliable as to not require extensive editing. Studies in event detection as described e.g. in the Aladdin Project are also expected to provide substantial support for advanced indexing and annotation in MMC.

**Summary**

This chapter has explored different avenues to enhance MMC for the purposes of AVT research. Until now, research on multimodality has been mostly restricted to face-to-face interaction, without much consideration of visual or acoustic elements in the surrounding situation. From the point of view of AVT and related research, however, it is necessary to consider a very large amount of visual and acoustic element of the situations represented in the objects to be translated. It has therefore been proposed to enhance the notion of multimodality (and MMC) in order to accommodate all translation-relevant audiovisual products. Another kind of enhancement would involve shaping parallel corpus architectures. Unfortunately, multimodality and parallelism dramatically increase the complexity of corpora in a number of respects. In particular, annotation schemes may quickly become unmanageable and query interfaces too complicated. Future multimodal corpora for translation purposes will probably evolve into two types: small corpora for scholarly research,
Multimodal corpora and AVT studies

and larger ones for more general purposes. Larger corpora will be dependent on substantial progress in automatic annotation, and probably will have to restrict their annotation schemes to automatically identifiable features.

Further reading


Related topics

12 Mediality and audiovisual translation
13 Spoken discourse and conversational interaction in audiovisual translation
17 Multimodality and audiovisual translation: cohesion in accessible films
20 Corpus-based audiovisual translation studies: ample room for development
32 Technologization of audiovisual translation

References


**Sitography**

Aladdin Project: http://multimedia.icsi.berkeley.edu/video-content-analysis/aurora-an-aladdin-project/ [last access 20 December 2017].

AntConc | Freeware corpus analysis toolkit for concordancing and text analysis: http://www.laurenceanthony.net/software/antconc/ [last access 20 December 2017].

ANVIL | Video annotation research tool: http://www.anvil-software.org/ [last access 20 December 2017].

British National Corpus (BNC) | http://www.natcorp.ox.ac.uk/ [last access 20 December 2017].
ConcGram | Phraseological search engine: https://benjamins.com/#catalog/software/clsl/main [last access 20 December 2017].
DeReKo | Das Deutsche Referenzkorpus: http://www1.ids-mannheim.de/kl/projekte/korpora/ [last access 20 December 2017].
DWDS | Das Wortschatzsystem zur deutschen Sprache in Geschichte und Gegenwart: https://www.dwds.de/ [last access 20 December 2017].
EXMARaLDA | Werkzeuge für mündliche Korpora: http://exmaralda.org/en/ [last access 20 December 2017].
PRAAT | Doing phonetics by computer: http://www.fon.hum.uva.nl/praat/ [last access 20 December 2017].
Sketch Engine | Corpus manager and text analysis software: https://www.sketchengine.co.uk/ [last access 20 December 2017].
XTrans | Multi-platform, multilingual, multi-channel transcription tool: https://www.ldc.upenn.edu/language-resources/tools/xtrans [last access 20 December 2017].