

The web as a disposable corpus for conference interpreters

Dr. Ilhem Bezzaoucha *

Institut de Traduction, Université Alger

ilhembazzaoucha@gmail.com

DOI:10.33705/1111-017-002-015

Received: 14/10/2024

Accepted: 26/12/2024

Published: 31/12/2024

*Corresponding Author

Citation :
Bezzaoucha, I. (2024).
The web as a disposable corpus for
conference interpreters
Maalim
I(2), 21-32

Abstract:

With the rapid advancement of technology, conference interpreters increasingly use the web as a disposable corpus to aid in their work. This paper examines how the web provides dynamic, accessible, and up-to-date linguistic data to support interpretation tasks. While it offers immediate access to relevant information, challenges such as the reliability of online content and ethical considerations arise. Through practical examples, the paper highlights how interpreters can integrate web resources into their workflow, enhancing their performance. It also addresses issues of data privacy and copyright compliance, advocating for the web's potential to support continuous professional development in the digital age.

Keywords: conference Interpreters, disposable Corpus, web resources conference documentation interpretation practice.

Maalim
© 2024 The Author(s).
Published by the High council of the Arabic
language.
This is an open access article
under the [CC BY license](#)



Résumé :

Avec l'avancée rapide de la technologie, les interprètes de conférence utilisent de plus en plus le web comme corpus jetable pour les aider dans leur travail. Cet article examine comment le web offre des données linguistiques dynamiques et accessibles pour soutenir les tâches d'interprétation. Bien qu'il fournisse un accès immédiat à des informations pertinentes, des défis tels que la fiabilité des contenus en ligne et les considérations éthiques se posent. Grâce à des exemples pratiques, l'article montre comment les interprètes peuvent intégrer ces ressources à leur flux de travail, tout en abordant les questions de confidentialité des données. Il souligne l'importance du web pour le développement professionnel continu dans un monde de plus en plus numérique.

Mot clés : Interprètes de conférence, Corpus jetables ; ressources Web, documentation de conférence, pratique d'interprétation.

شبكة الإنترنت باعتبارها مجموعة يمكن التلخيص منها لمترجمي المؤتمرات

الملخص:

مع التقدم السريع للتكنولوجيا، يستخدم مترجمو المؤتمرات الفوريون (الويب) بشكل متزايد كمجموعة يمكن التلخيص منها للمساعدة في عملهم. تبحث هذه الورقة في كيفية توفير (الويب) لبيانات لغوية ديناميكية وسهلة الوصول وحديثة لدعم مهام الترجمة الفورية. في حين أنها توفر إمكانية الوصول الفوري إلى المعلومات ذات الصلة، إلا أن هناك تحديات مثل موثوقية المحتوى عبر الإنترنت والاعتبارات الأخلاقية. ومن خلال الأمثلة العملية، تسلط هذه الورقة البحثية الضوء على كيفية دمج المترجمين الفوريين لموارد الويب في سير عملهم، مما يعزز أداءهم. كما تتناول الورقة أيضاً قضايا خصوصية البيانات والامتثال لحقوق الطبع والنشر، ويدافع عن إمكانات (الويب) لدعم التطوير المهني المستمر في العصر الرقمي.

الكلمات المفتاحية: مترجمو المؤتمرات، المتون القابلة للتصرف؛ وثائق مؤتمرات موارد الويب، وممارسة الترجمة الفورية.

1. Introduction:

Language users are becoming increasingly aware of the tremendous advantages of using the web as a source of fairly reliable linguistic information. A simple task as spell-checking is often performed by professionals and beginners alike. The web has become a one of a kind free marketplace to shop for linguistic validation.

We will argue that small specialized corpora from a source and a target language can become additional helpful resources of specialized language tracking for interpreters. Translators and interpreters alike have been known for retrieving ad-hoc corpora for specific work assignments from the huge and exponential mass of textual data that circulate freely in the web. The web however, is home to unstructured, tangled information freely accessible to the public. This paper will investigate

ways for conference interpreters to retrieve additional targeted textual information from the gigantic information supplier that is the World Wide Web.

1- Conference documentation

The internet has progressively evolved into a catalyst for the acquisition and dissemination of linguistic knowledge. Traditional dictionaries, while valuable, often struggle to keep pace with the constant influx of new terms and expressions. This inability to promptly address the diverse linguistic needs of users underscores the growing significance of online resources in meeting immediate language-related requirements.

In the realm of interpretation, practitioners are acutely attuned to the nuances of language usage. As Gile (2009) elucidates, interpreters engaged in highly specialized conferences must not only possess linguistic proficiency but also gather pertinent linguistic and contextual information to ensure the efficacy of their interpretation tasks. This underscores the pivotal role of interpreters in facilitating effective communication and knowledge exchange in diverse settings.

It is imperative for event organizers to acknowledge the pivotal role of interpreters in shaping the success of an event. The quality of interpretation significantly influences the overall experience of participants, underscoring the need for organizers to prioritize the selection of skilled interpreters to ensure seamless communication and comprehension among attendees.

Additionally, the creation and maintenance of glossaries have long been regarded as integral components of an interpreter's toolkit. The International Association of Conference Interpreters (AIIC) acknowledges the widespread practice among interpreters of compiling glossaries to aid in terminology management and preparation for assignments. These glossaries serve as invaluable reference materials, enabling interpreters to access and verify specialized terminology promptly during interpretation sessions. These glossaries are the product of targeted preparation

Conference documentation, however, is not always made available:

"There are several reasons why conference documents are not always made available to interpreters in advance: papers are often finalized at the last moment, speakers are not always made aware of the interpreters' needs, they may not wish to disclose the content of their papers in advance, they may consider their papers confidential and are afraid of security breaches." (Gile 2009)

2. Corpora in interpreting studies

Interpreters frequently encounter challenges akin to those faced by translators, as they constantly seek optimal linguistic alternatives in the target language. They rely on web texts and

utilize popular search engines to retrieve real-life examples of words and phrases in contextual settings.

The term "corpus" has been widely used in Translation Studies to denote a collection of texts that translators and researchers analyze for specific research objectives. When translators handle a translation task, they seek out various types of texts to extract suitable terms and phrases, thereby enhancing the naturalness of target texts. Similarly, interpreters can construct their own parallel corpora in their language pairs by leveraging web pages offering transcribed versions of speeches delivered at official events. However, it is important to note that the web is an anarchic and chaotic source of information, requiring interpreters to navigate carefully.

Parallel and comparable corpora are the predominant types of corpora encountered in translation contexts. These corpora facilitate linguistic analysis by providing translators and interpreters with valuable insights into language usage and patterns across different language pairs.

One of the most common applications of corpus linguistic data is the extraction of a "language in use" model. This model comprises a weighted combination of words that reflect their relationships to each other, their prevalence in specific domain-specific fields, and the strength of the network they form. By analyzing such linguistic patterns, interpreters can gain deeper insights into the nuances of language usage within specific contexts, enabling them to produce more accurate and contextually appropriate interpretations.

To address the aforementioned challenges and gain a comprehensive understanding of the current preparation practices of conference interpreters, it is imperative to conduct rigorous research. This research should explore the strategies and techniques employed by interpreters to perform effectively under tight deadlines. By investigating the utilization of corpora and other linguistic resources in interpreter training and practice, scholars can identify best practices and areas for improvement in interpreter preparation and performance.

3. the web as corpus

The utilization of the web as a corpus represents a significant aspect of contemporary interpreting practice, offering interpreters access to a vast repository of linguistic data. According to the International Association of Conference Interpreters (AIC), interpreters can draw upon three primary sources of information to prepare for upcoming events: conference documents, the World Wide Web, and other background information and terminology resources provided during pre-conference briefings (AIC website). However, the web is characterized by its sprawling unregulated nature, as noted by Varantola (2003), who describes it as a "huge, unregulated, and messy source of

disposable corpora."Despite its chaotic structure, the web harbors an abundance of language data, unparalleled in richness and accessibility.

Kilgarriff (2003) describes the web as a "dirty corpus,"acknowledging its imperfections while highlighting the frequency of expected usage. Zanettin (2012) further underscores the disparities between the web and traditional corpora, emphasizing differences in the accessibility, display, and interpretation of linguistic data. Navigating the web to select ad hoc corpora is essential for interpreters, as stand-alone documentation sources often fail to encompass all the terminological needs of an interpreting assignment. As Chomsky elucidates, language is a dynamic process of creation, where principles of generation are fixed, but their utilization is infinitely varied.

It is argued that a diverse array of tools enhances the performance of language professionals. Disposable corpora derived from the web can heighten interpreters' awareness of terminological phraseological, and structural nuances, ultimately leading to greater client satisfaction. Jensen Moustén, and Laursen (2016) highlight a shift in translation scholarship towards smaller corpora of specialized text, which better capture the traits of specific genres or domains. Term extraction proves beneficial to interpreters only when accompanied by a broad understanding of actual term usage in authentic texts. Comparable corpora usage can mitigate interference from the source language, a common occurrence in parallel texts.

A clear distinction exists between interpreting competence and actual performance, with the latter heavily reliant on reference sources to retrieve real examples of words and phrases in contextual settings. By leveraging the web as a corpus, interpreters can enrich their linguistic repertoire, refine their terminological accuracy, and ultimately enhance their performance in interpreting assignments.

4. Main concerns and applications

In navigating the vast and unstructured landscape of the World Wide Web, interpreters face a formidable challenge in harvesting additional targeted information amidst the anarchic and messy nature of online sources. The internet, characterized by its sprawling and unregulated content poses significant hurdles for interpreters seeking to extract relevant linguistic data for their interpreting assignments. However, despite these challenges, interpreters employ various strategies to sift through the chaos and retrieve valuable information to support their interpretation tasks.

One approach interpreters utilize to navigate the web effectively is the utilization of advanced search techniques and tools. By employing specialized search operators, interpreters can refine their search queries and narrow down results to focus on specific domains or topics relevant to their

interpreting assignments. Additionally, interpreters may leverage the capabilities of advanced search engines and online databases designed to index and organize web content more efficiently. These tools enable interpreters to access targeted information more effectively amidst the vast expanse of online data.

Furthermore, interpreters may rely on curated collections of online resources and databases tailored to their language pairs and subject matter expertise. These curated collections, often maintained by professional associations or language service providers, provide interpreters with access to pre-selected and vetted resources relevant to their interpreting assignments. By leveraging these curated collections, interpreters can streamline their information retrieval process and access high-quality linguistic data more efficiently.

In addition to harnessing the power of advanced search techniques and curated collections interpreters may also employ corpus-based approaches to gather targeted information from the web. Corpus linguistics, a field that focuses on the analysis of large collections of linguistic data offers interpreters valuable insights into language usage patterns and conventions across different domains and contexts. By compiling and analyzing corpora of web texts relevant to their interpreting assignments, interpreters can identify recurring linguistic patterns, specialized terminology, and context-specific language usage, enhancing their ability to render accurate and contextually appropriate interpretations.

However, one might question why interpreters would resort to corpus data in interpreting assignments when glossaries and databases are available as specialized lexical resources. While glossaries and databases provide valuable context-independent information on word usage interpreters often encounter lexical problems that require context-dependent solutions. As noted by Varantola (2003), interpreters have idiosyncratic and sophisticated needs, necessitating the ability to make informed decisions based on cumulative evidence gathered from various sources during information retrieval operations.

Interpreters may turn to corpus data as a means of accessing real-life language usage examples and contextually relevant linguistic data to inform their interpretation decisions. Unlike glossaries and databases, which offer static and pre-defined sets of terms and definitions, corpora provide interpreters with dynamic and contextually rich linguistic data extracted from authentic texts. This allows interpreters to glean insights into language usage patterns, collocations, and idiomatic expressions, enabling them to produce more accurate and natural-sounding interpretations.

Moreover, interpreters may utilize corpus data to overcome challenges associated with interference from source languages in parallel translated texts. By analyzing non-translated texts and compiling ad-hoc corpora specific to their interpreting assignments, interpreters can minimize the risk of interference and enhance the quality and fidelity of their interpretations.

The concept of compiling ad-hoc corpora, also known as corpora "on the fly," for specific interpreting events is a common approach utilized by interpreters to enhance the productivity and quality of their work. This method involves assembling a collection of relevant texts and linguistic data tailored to the specific requirements of an interpreting assignment.

[Hyunjoo Park](#) (2010) discusses this practice in the context of interpreting, highlighting its effectiveness in providing interpreters with access to real-life language usage examples and contextually relevant linguistic data. By compiling ad-hoc corpora specific to their interpreting assignments, interpreters can gather valuable insights into language usage patterns, specialized terminology, and context-specific language usage, ultimately enabling them to produce more accurate and contextually appropriate interpretations.

Furthermore, McEnery and Wilson (2001) emphasize the importance of sampling and representativeness in compiling corpora for interpreting purposes. Sampling refers to the process of selecting a representative subset of texts or linguistic data that accurately reflects the language usage and characteristics of the target domain or context. Representativeness ensures that the compiled corpus adequately captures the diversity and variability of language usage within the target domain or context, thereby enhancing its utility and applicability for interpreting purposes.

The benefits of compiling ad-hoc corpora for interpreting purposes are manifold. Firstly digital versions of texts are often freely available for download and reading, facilitating the efficient compilation and accessibility of linguistic data. This accessibility enables interpreters to quickly gather relevant texts and linguistic resources tailored to their interpreting assignments, thereby streamlining the preparation process.

Secondly, ad-hoc corpora enable interpreters to engage in real-time study of current events allowing them to track unfolding language usage and developments relevant to their interpreting assignments. This real-time access to linguistic data ensures that interpreters remain up-to-date with the latest language trends and developments, enhancing the accuracy and relevance of their interpretations.

5. Corpus-analysis tools

Corpus-analysis tools play a crucial role in the arsenal of language professionals, particularly interpreters, seeking to enrich their linguistic resources and enhance their performance. These tools facilitate the compilation and analysis of corpora, enabling interpreters to access and explore vast collections of linguistic data for their interpreting assignments.

Web-based corpora offer interpreters the opportunity to manually construct corpora through queries and downloads, albeit through a time-consuming process. Interpreters can utilize concordancing software to navigate these corpora, extracting valuable insights and linguistic patterns. However, this approach is often deemed laborious, as the resulting corpora are typically intended for single-use purposes.

Alternatively, interpreters can leverage automated tools such as Sketchengine to compile corpora efficiently. This software automates the process of corpus compilation, allowing interpreters to focus their efforts on analyzing linguistic data rather than constructing corpora manually. By utilizing concordancing software in conjunction with automatically compiled corpora, interpreters can glean valuable linguistic information with greater ease and efficiency.

Interpreters are keen to enrich their own glossaries for future reference, but they are cautious of unfamiliar or emerging language usage that may evade their recognition. As such, interpreters prioritize the gathering of long stretches of language rather than isolated lexical items. Corpora consist of texts of varying lengths, enabling interpreters to contextualize their analyses and gain a deeper understanding of language usage patterns across different contexts and domains.

Ad hoc knowledge acquisition for interpreting assignments typically occurs prior to the relevant conference, in contrast to translation, where it often takes place during the translation process (Gile, 2009: 131). This distinction underscores the proactive approach adopted by interpreters in preparing for their assignments, as they strive to familiarize themselves with the language and subject matter in advance.

Linguee, a prominent web interface, exemplifies the convergence of corpus-driven analysis and concordancing functionalities. In addition to serving as a multilingual online dictionary Linguee functions as a concordancer, enabling interpreters to locate potential equivalents for specific terms and phrases within limited contextual information. With access to over one billion translated texts sourced from professionally translated websites, Linguee provides interpreters with invaluable insights into translation solutions employed by their peers, facilitating knowledge sharing and professional development.

In recent years, the BootCaT toolkit has gained prominence in various experiments aimed at bootstrapping corpora from the web to extract linguistic information such as terms or collocations. Notable studies utilizing BootCaT include those by Baroni and Bernardini (2004 and Sharoff (paper below). Our implementation of multi-word term extraction bears some resemblance to the method proposed by Baroni and Bernardini (2004).

The fundamental BootCaT procedure is straightforward, accomplishing two main tasks: 1) constructing a corpus of specialized texts sourced from the web, and 2) extracting relevant terminology from the compiled corpus. BootCaT adopts a common approach in terminology extraction and corpus comparison, comparing frequencies in specialized and reference corpora to identify terms characteristic of the former. This iterative algorithm bootstraps corpora from the web and initially extracts unigram terms. Subsequently, it proceeds to extract multi-word terms based on the downloaded corpus and the unigram term list obtained earlier.

The bootstrapping process, facilitated through the Google search engine, commences with a small list of seeds assumed to be representative of the domain. These seed terms are randomly combined to form Google query strings, with each combination serving as a query to retrieve the top *n* pages (HTML, PDF, and doc files). The retrieved pages are then formatted as text, and unigram terms are extracted from this corpus by comparing their frequency of occurrence with that in a reference corpus.

6. How a corpora are disposable for conference interpreters?

Disposable corpora represent a flexible and transient resource for conference interpreters tailored to the unique demands of specific interpreting assignments or events. Unlike permanent corpora that endure for long-term linguistic analysis or translation, these corpora are assembled and utilized for a temporary purpose, often becoming obsolete once the interpreting task concludes. This essay delves into the concept of disposable corpora for conference interpreters exploring their utility, creation, and eventual disposal.

In the realm of conference interpreting, disposable corpora serve as indispensable aids in the preparation and execution of interpreting assignments. These corpora typically comprise a curated selection of texts, transcripts, speeches, or presentations pertinent to the subject matter and context of the conference or event being interpreted. Sources for disposable corpora can vary widely, ranging from conference materials and official documents to online resources and specialized databases. The key characteristic of disposable corpora is their relevance and applicability to a specific interpreting task or event.

One of the primary functions of disposable corpora is to provide interpreters with access to real-world language usage examples and contextually relevant linguistic data specific to the event they are interpreting. By immersing themselves in the terminology, style, and subject matter of the conference, interpreters can familiarize themselves with the nuances of the discourse, enabling them to deliver accurate, contextually appropriate interpretations. Disposable corpora thus serve as invaluable tools in supporting interpreters' linguistic proficiency and ensuring the quality of their interpretations.

The process of creating disposable corpora involves careful curation and selection of relevant texts and linguistic data tailored to the specific requirements of the interpreting assignment. Interpreters may employ various strategies to compile disposable corpora, including gathering conference materials, official documents, and supplementary resources from online sources. Additionally, interpreters may leverage specialized databases and linguistic tools to enhance the comprehensiveness and accuracy of their corpora. The goal is to create a comprehensive and representative corpus that reflects the linguistic landscape of the conference or event, enabling interpreters to prepare effectively for their interpreting tasks.

Once the interpreting assignment is completed, disposable corpora may no longer serve a practical purpose and may be considered disposable. While some interpreters may choose to archive their corpora for future reference or analysis, others may opt to discard them, freeing up storage space and resources for subsequent projects. The disposability of these corpora reflects their transient nature, designed to meet the immediate needs of a specific interpreting task or event. Despite their temporary nature, disposable corpora play a vital role in supporting conference interpreters throughout the interpreting process. By providing access to relevant linguistic data and contextually rich language usage examples, these corpora enable interpreters to prepare thoroughly and deliver accurate interpretations. Moreover, disposable corpora facilitate adaptability and flexibility, allowing interpreters to tailor their linguistic resources to the unique requirements of each interpreting assignment.

Conclusion

Conference interpreters often face the challenge of rendering complex linguistic and structural choices from the source language into the target language accurately and effectively. In this endeavor, interpreters can leverage readily-made interpreted translated texts as valuable resources to enhance their interpretation process. These interpreted translations serve as rich repositories of

linguistic and structural choices made by experienced interpreters, providing insights and guidance for interpreters encountering similar challenges in their own interpreting assignments.

By analyzing interpreted translated texts, interpreters can glean valuable information about how to handle specific linguistic nuances, idiomatic expressions, and stylistic conventions present in the source language. These texts offer concrete examples of successful interpretation strategies employed by experienced professionals, allowing interpreters to learn from their peers and refine their own interpretation techniques. Additionally, interpreted translated texts provide interpreters with a wealth of context-specific vocabulary and terminology, enabling them to expand their linguistic repertoire and improve the accuracy and fluency of their interpretations.

Furthermore, corpora serve as invaluable supplements to conventional resources such as dictionaries and glossaries, which may not always cater to the varied contextual needs of task-oriented users like conference interpreters. While dictionaries and glossaries offer definitions and translations of individual words and phrases, they may lack the depth and specificity required to address the complex linguistic challenges encountered in conference interpreting. In contrast corpora provide interpreters with access to authentic language usage examples drawn from real-life contexts, allowing them to better understand the nuances of language and make informed interpretation decisions.

Moreover, corpora offer interpreters the opportunity to explore language patterns collocations, and discourse structures prevalent in specific domains or subject areas. By analyzing corpora compiled from texts relevant to their interpreting assignments, interpreters can identify recurring linguistic patterns and conventions, enabling them to produce more accurate and contextually appropriate interpretations. This deeper understanding of language usage allows interpreters to convey the intended meaning and tone of the source message faithfully, thereby enhancing communication and comprehension in multilingual settings.

References

1. Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the Special Issue on the Web as Corpus. September 2003. Computational Linguistics 29(3).
2. Baroni, M and S. Bernardini 2004. BootCaT: Bootstrapping corpora and terms from the web. Proceedings of LREC 2004, Lisbon.
3. Baroni M. and A. Kilgarriff (2006) Large linguistically processed web corpora for multiple languages. Proc. EACL. Trento, Italy.

4. Fantinuoli C. 2006. Specialized corpora from the Web and term extraction for simultaneous interpreters. In M. Baroni and S. Bernardini, editors, *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna.
5. Gile, D. (2009) *Basic Concepts and Models for Interpreter and Translator Training: Revised edition* DOI: [10.1075/btl.8](https://doi.org/10.1075/btl.8).
6. [Hyunjoo Park](#) (2010): [FORUM. Revue internationale d'interprétation et de traduction/ International Journal of Interpretation and Translation, Volume 8, Issue 1](#), Jan 2010, p. 267.
7. Kilgariff, A, P. Rychlý, P. Smrz and D. Tugwell. 2004. The Sketch Engine. Proc Euralex. Lorient.
8. Moustén, and Laursen (2016) *Specialized Languages-the alter ego of any research field* https://www.researchgate.net/publication/309284247_Specialized_Languages-the_alter_ego_of_any_research_field.retrived 08/08/2024
9. McEnery, T. and Wilson, A. (2001) *Corpus Linguistics An Introduction*. 2nd Edition Edinburgh University Press, Edinburgh.
10. Sharoff, S. 2005. Creating general-purpose corpora using automated search engine queries <http://wackybook.sslmit.unibo.it/pdfs/sharoff.pdf>. Retrieved 08/08/204.
11. Varantola, K. (2003) Translators and disposable corpora. In F. Zanettin, S. Bernardini and D. Stewart (eds) *Corpora in Translator Education*. Manchester: St Jerome, pp. 55-70.
12. Zanettin, F (2012), *Translation-Driven Corpora (Translation Practices Explained 14)*. Manchester: St. Jerome Publishing.