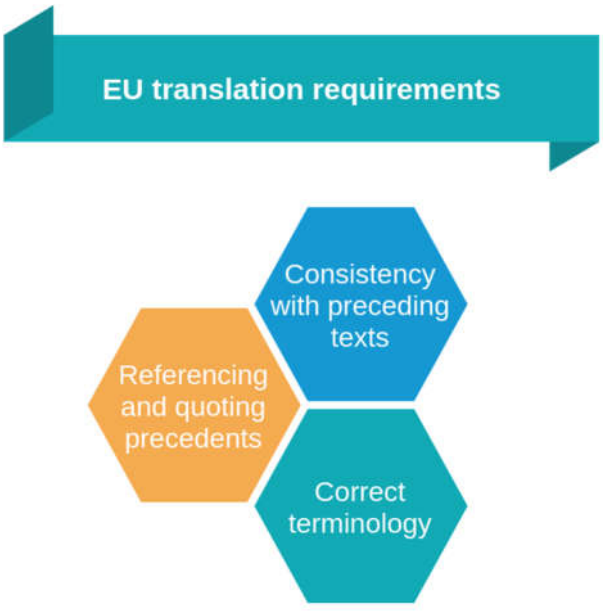
Complying with the quality requirements of EU institutional translation: an overview of terminology and concordance search tools

Preamble

In this paper we will introduce the quality requirements of translation projects delivered to the bodies of the European Union, and enumerate the challenges of fully complying with these standards. We categorize and review some of the most important tools which are currently - either publicly or internally - available to linguists for performing terminology research in EU-related texts. Next, we introduce a solution which offers fast and comprehensive online concordance search over the EU legal corpus and the IATE termbase. We show the advantages this tool combines through its EU-translation specific features. Finally, we illustrate the benefits of Juremy.com by analysing typical usage patterns.

I. What are the most important challenges translators face in EU institutional translation projects?

Translation for the Bodies of the European Union requires high quality standards, such as *consistency* with preceding EU texts, exact *referencing* of, and correct quotation from published documents, and consistent use of *correct terminology*.
To meet these standards continuously and thus maintain a high-ranking position among contractors, translators need to spend a significant amount of time on terminology research and referencing.



For example, the Tender Specifications of the European Commission (TRAD19) sets out among others that

- "references to and explicit or implicit quotes from published documents must be checked and quoted correctly", or

- “correct terminology must be used consistently throughout the text in line with the relevant domain, reference documents and appropriate naming conventions”.

Similar requirements are found in the Tendering Specifications of the Court of Justice of the European Union, which ask contractors to ensure “strict use of the legal terminology used in the reference documents (source and target languages); rigorous citation of the relevant legislative and/or judicial texts, and use of the necessary legal databases (of the European Union and national)”.

The Tendering specifications of the European Parliament also include similar standards under the chapter “Quality requirements”: The Contractor must ensure that “all references to documents already published or any reference material, including the terminology of the reference material quoted, are consulted and used correctly”.

No wonder that compliance with the above consistency and referencing standards of EU institutional translation requires extensive terminology research from the linguists’ part. Due to the consistency requirement for example, translators must research the Union's corpus vigilantly to discover precedents of a given term already translated into the target language. Furthermore, to support their choice of a specific term, they need to provide references to already published documents by inserting a relevant hyperlink or document number.

A practical example: when translating “social security” from English to Hungarian, you would normally use the term “társadalombiztosítás”. However, the Treaty on the functioning of the EU already features it as “szociális biztonság”, which sets a precedent. The IATE database labels the unconventional “szociális biztonság” as “Preferred” as well - contrary to the common national grammatical usage. Due to the consistency requirement, you must perform terminology research vigilantly to discover such precedents.

This is all the more important given that the EU institutions regularly assess the overall performance of framework contracts by checking the quality of the delivered pages and grading them from “very good” to “unacceptable”. The failure to satisfy the quality requirements may lead to worse ranking during the evaluation process vis-a-vis other contractors, in more severe cases penalties, or in the worst case even termination of the framework contract.

II. The main technical means available for performing terminology research on the EU corpus

In this chapter, we will go through the currently available tooling closely related to EU translation. These tools can be classified along various dimensions.

Online or local: Online tools need less management - they don’t need installation, but rely on internet connectivity for operation. Local tools can integrate with other software and workflows directly, but need installation and management of offline data sources. There’s also crossover in-between, where a local tool accesses online resources as well (this access can be instant / synchronous, or batched / asynchronous).

The provided resource can be Document Repository (for lookup of relevant monolingual documents), Termbase (for retrieving bilingual terminology like phrases), Translation Memory (for finding longer segments of existing bilingual corpora or previously translated documents), and more recently Machine Translation (for pre-translating documents for later verification and post-editing).

Both tools and resources can either be publicly available, or developed for private (institutional) use.

Table 1: Examples of publicly available resources

publicly available	ONLINE	
Document repository	EUR-Lex, Curia	custom collection
Termbase	IATE	IATE data dump
Translation Memory	Linguee, Glosbe	DGT-TM
Machine Translation	DeepL; eTranslation (EC)	-

Table 2: Examples of internally available or private resources

internally available	ONLINE	OFFLINE (+ CAT tool)
Termbase	EU-internal SDL / IATE integration; IATE search for a set of entries (batch operations	private terminology databases
Translation Memory	Euramis for EUR-Lex	private TM collections, custom alignments, per-job generated extracts
Machine Translation	eTranslation - an online MT service provided by the European Commission	-

- The official online EUR-Lex search site provides the official and most comprehensive access to EU legal documents, translated in the highest quality to the 24 official EU languages. Due to its purpose however, it is designed for document search, and not for the linguists' use case. Using Eur-Lex for bilingual concordance search is thus a lengthy process, and users often face misaligned paragraphs within bilingual text display.

- *IATE* (Interactive Terminology for Europe) is the EU's concept-oriented multilingual termbase, used for the collection, dissemination and management of **EU-specific**

terminology. The data in IATE is mostly entered by translators and terminologists working in the language services of EU institutions. IATE contains nearly 1 million entries, which are reviewed by terminologists, with additional metadata such as domain classification based on the *EuroVoc thesaurus*, the evaluation of the term (preferred, deprecated, obsolete, proposed or admitted) or reliability code (1 to 4 stars).
IATE is an annotated and trusted source, but has limited coverage. Also, the approximately 1 million entries do not include all language variations of a given term.

Internal translators of the bodies of the EU also have access to document-specific termbase retrieval in IATE.

- *DGT-Translation Memory* is a free collection of aligned segments, mainly from texts of EU legislation (sector 3 of the Celex classification) translated in all EU official languages. Users need to follow technical instructions on where to download the DGT-TM and how to produce bilingual aligned corpora in the preferred language pair, using a special extraction tool. The number of documents covered by the database is around 40,000. However, the time coverage of 2004 - 2020 is smaller than that of EUR-Lex, and the corpus is limited to partial coverage of EU legislation. The DGT-TM also covers some documents in sector 2 (International agreements) but does not contain case law (judgments, opinions of the advocate general, etc.). The refresh happens on a yearly basis.

- *Online concordance search* has been popularized by services like Glosbe and Linguee. They cover a wide range of corpora, which at the same time implies they can't target our focus use case of EU legal and professional translation. The exact source of data is usually unknown. The completeness of the import and the update frequency is also unknown. They usually also lack precise reference to the source document. The hit segment contains links to the pdf version of the monolingual Official Journal, making bilingual verification inconvenient.

- *CAT Tools* are offline, or locally managed tools incorporating both search and editing capability. Their speed and coverage depends on the quantity and quality of data loaded into them, and tasks the maintainer with content upgrades. To be used more specifically to EU institutional translation purposes, CAT Tools might be customized for specialized use via plugins. For example, IATE has developed a plugin for Trados Studio, which has an asynchronous Term Recognition module and a near-synchronous Term Provider feature which is segment based. However, these features are dedicated to internal users, and not available to external users working outside the EU institutions.

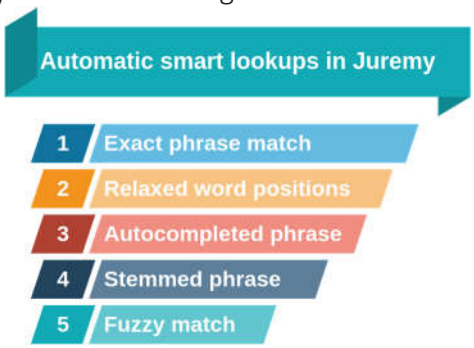
- Another tool which is only available to EU institutions' staff is *Euramis* (**Eur**opean **A**dvanced **M**ultilingual **I**nformation **S**ystem). Euramis is a system of databases or memories (and associated tools) fed by the various institutions of the EU. It acts as a multilingual, multidirectional repository of equivalent phrases ("segments") belonging to official EU documents allowing their re-use in translation in all European Institutions.

III. How are the advantages of the above tools combined by a publicly available solution?

The EU-translation focused advantages and features which the above mentioned tools offer are combined in Juremy.com, a fast and comprehensive online concordance search tool operating over the EU legal corpus and the IATE termbase. This tool lifts the burden on EU translators of installing a term extraction software, and of acquiring, processing, making searchable and updating the database serving as a termbase and a translation memory for EU institutional translation projects.
Juremy's database covers 278,000 documents in sectors 1, 3, 5, 6 (Treaties, Legislation, Preparatory documents, EU case-law), totalling more than 700 million segments over 24 languages.

Juremy has a larger database coverage than the above tools (except Eur-Lex), which is regularly updated (every month). We perform customized text alignment on the language variants, so the user does not need a separate software to align the different segments of the documents in the preferred language pair (this alignment is usually performed by CAT Tools).

In addition to exact phrase matching, Juremy automatically executes the following smart searches: fuzzy search, relaxed word positions–, autocomplete phrase–, stemmed phrase– and optionally, unaccented exact phrase search. The results are sorted by weighting shortness of match and fundamentality of the containing document.



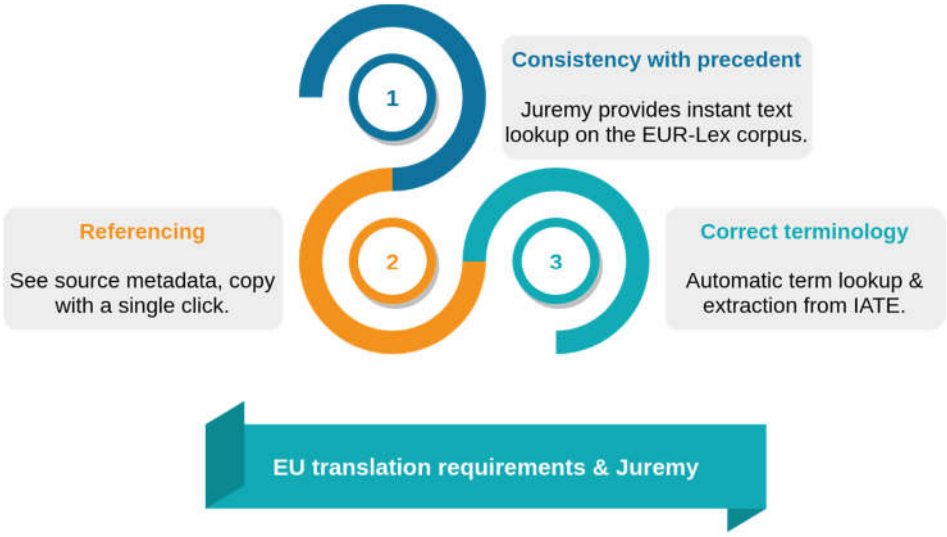
The way Juremy displays results supports the EU translators' workflows. The results feature precise source attribution and linking the source material, displaying relevant metadata, and providing domain-specific document filters.
It is ideal for freelancers or agencies who do not manage a comparably extensive local translation memory themselves, or lack tooling customizations focusing on the EU translation

requirements. It has no maintenance burden, and is periodically updated with data and new features as well.

Since its creation, we gradually added the current functionalities to our program, so in its present form our tool

- periodically indexes the documents published on EUR-Lex,
- covers Celex sectors 1, 3, 5 and 6 beyond DGT-TM,
- includes IATE entries too, and optionally extracts all exactly matching bilingual IATE terms within the query phrase - either multi-word or single-word,
- performs concordance search with bilingual paragraph-level results in less than a second,
- corrects the often misaligned paragraphs in EUR-Lex documents,
- automatically performs smart fuzzy searches besides exact phrase search,
- can search using any language pair over the 24 official EU languages,
- displays relevant document metadata, such as Celex number and Eurovoc topics to streamline the EU terminology research and referencing workflow,
- makes citation easy by one-click copy of the bilingual document link to clipboard.

As a result, you can immediately see the most relevant matches for a term or phrase in the EU corpus, without spending valuable time on browsing and clicking in various databases. Ultimately, our goal is to help eliminate the repetitive and laborious tasks from terminology research, and to display the most relevant matches to a given query through intelligent search and filtering. So in the end you only need to creatively synthesize the information surfaced, and you can spend your time saved by Juremy on more in-depth research or completing more projects within a given timeframe.



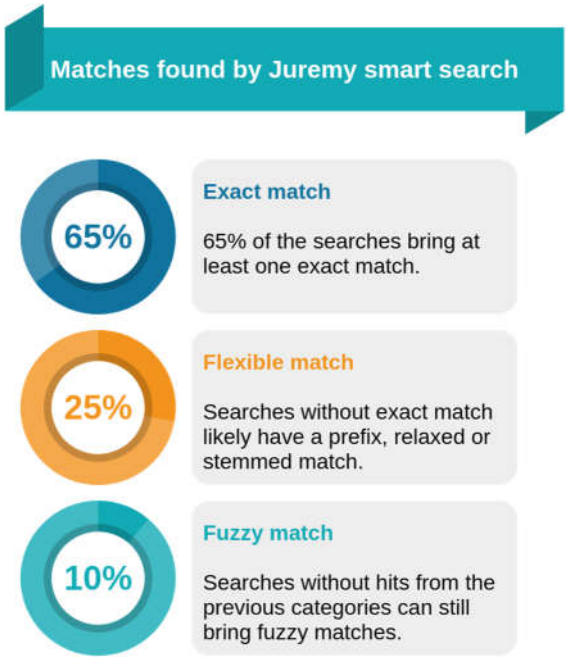
IV. How much time can be saved by speeding up the terminology search process?

As referred to above, it is the translator's task to diligently track down and annotate prior art of suspected phrases. As the volume of EU case-law constantly grows, and prior art accumulates, terminology research will consume more and more effort from the translators' side in order to comply with the growing requirements of coherence with other EU legal acts. To apply due diligence, translators annotate matched phrases with a match reference, most

often the document's Celex number. The go-to service for looking up matches so far is Eur-Lex. But the Eur-Lex interface was not designed for the translation use-case, rather topic research. For a term query, the user is presented with a list of documents. Manually opening these documents and again searching for the terms and then repeating this procedure until the perfect match of a term is found, especially in case of complicated expressions is a tedious task, which can easily take more than 5 minutes per query.

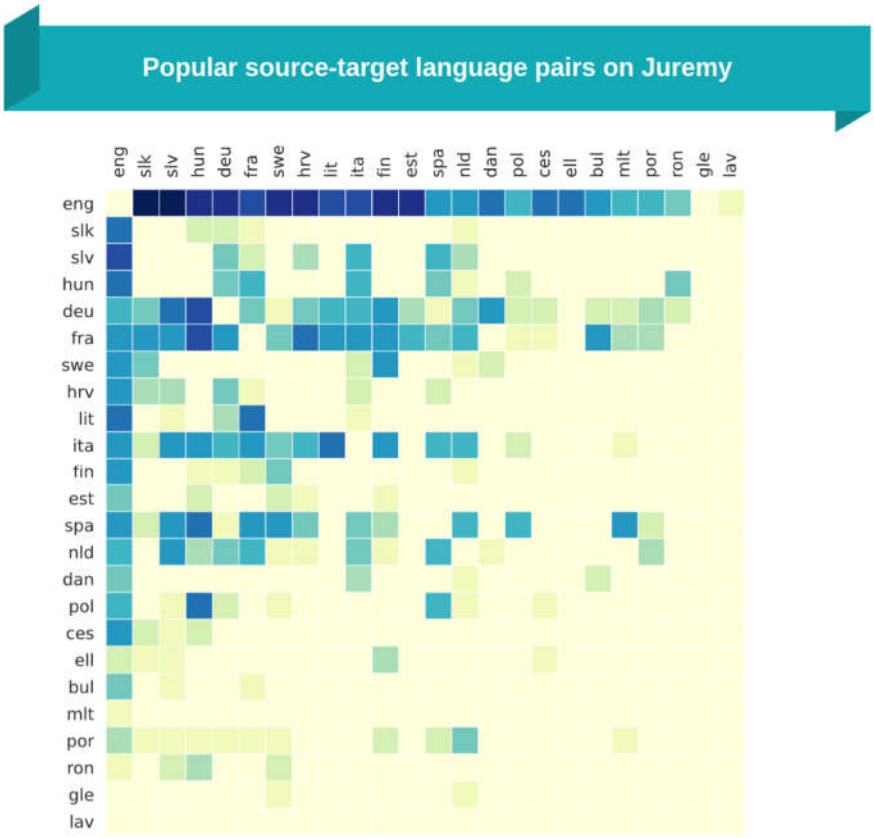
Distribution of smart search hit types

On the chart below we illustrated the coverage of matches found by Juremy smart search. For 65% of searches, there exists an exact hit. In the rest 35%, 25% has at least one of the relaxed position/ auto-completed/ stemmed phrase hits. The final 10% is only covered by the fuzzy hit.



Most frequently used language pairs in Juremy

On this chart we visualized the most frequently used language pairs in Juremy according to the most recent usage trends. It can be seen that the most common source language is English, followed by French and German. The most popular target languages are Central-European and Nordic ones.



Time saved by Juremy search

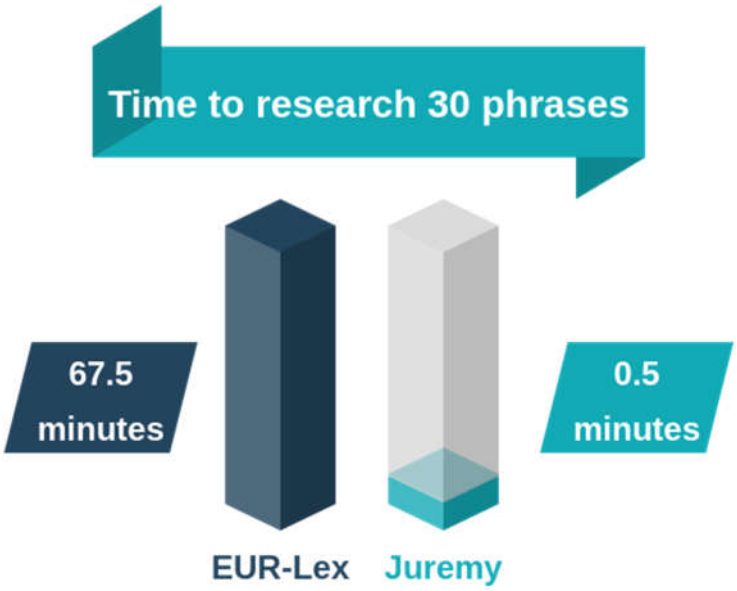
According to our analysis of Juremy usage, a typical subscriber performs 15 queries per day, while use during intense work ranges from 30 to 70 queries per day. Surfacing the exact matches takes Juremy less than 1 second in 95% of the cases.

Where an exact hit exists, we estimate that an average Eur-Lex search would take less than a minute, approx. 45 seconds (from the entry of the query in the search bar until the hit result is found in the bilingual text). In case of fuzzy matches however, in Eur-Lex the user has to apply different logic operators (AND, OR, " " for exact phrase search, * to replace characters, ? to replace a single character). Particularly in case of longer queries, it is unrealistic to attempt all possible logic operators within a phrase until a relevant hit is surfaced. Juremy performs these smart searches automatically.

If we do have any hit documents listed at all, finding the correct translation of a given phrase might require the opening and studying of several documents, also depending on the length of the query phrase. According to our experience, this task takes 5 minutes per query on average.

Aggregating the above results, we come to the following conclusion:
On an average day, 30 queries performed by a user to find the appropriate translation of a term takes 30 seconds (net) of a translator's time.

If these queries are performed on Eur-Lex, 65 % of them (~20) would display an exact hit which takes appr. 15 minutes. The rest of the queries (~10) would take 5 minutes of terminology search on average, i.e. 52.5 minutes. Consequently, the 30 queries on **Eur-lex takes up 67.5 minutes per day of a translator's time, which is 67 minutes more than the work spent with Juremy.** In case of the outlined scenario, Juremy saves a translator more than 1 hour of terminology research/day.



If you want to know Juremy better or you have any questions, you can read about all functionalities in detail in our User Manual, or feel free to contact us anytime at *info@juremy.com*.

juremy.com