AL-FARABI KAZAKH NATIONAL UNIVERSITY

Department of International Relations

Chair of Diplomatic Translation

**Translation business in the field of international and legal relations**

**“Translation of Scientific and Technical Documents”**

2024-2025 academic year, fall semester

**Lecture 7**

**Module 2: Resources for scientific and technical translation**

**Lecture 7: Researching TL terms**

Assuming we want to translate our text into Italian, it would be use- ful to compile an Italian corpus for this subject domain and to extract term candidates from that corpus, to help us to identify possible cor- responding terms for those we encountered in the ST. A small corpus of texts from the same genre or on the same topic, compiled relatively quickly, for the purposes of researching the language and termino- logical patterns of a subject domain, is known as an **ad-hoc corpus**, but may alternatively be referred to as a DIY (do-it-yourself) corpus or a disposable corpus. A specialized *ad-hoc* corpus can be much smaller than corpora that are used to gain a more general view of linguistic behaviour (Sinclair 2005).

Sketch Engine has a built-in tool called WebBootCaT that enables us to compile an *ad-hoc* corpus from the web using some key terms which are known as seeds (perhaps based on the principle of cloud seeding). If you do not have access to Sketch Engine you can down- load and install the BootCaT software as a stand-alone tool, free of charge, at bootcat.sslmit.unibo.it and use it to compile your corpus (see Kilgarriff 2013). ‘BootCaT’ stands for ‘bootstrapping corpus

and terms’.
As seeds, we could start with Italian terms we already know to be

relevant, like *sicurezza nucleare*, *materiale nucleare*, *combustibile nucleare*, *reattore nucleare*. With these terms as seeds and using WebBootCaT, Sketch Engine compiles a corpus of 53 files and 169,674 words at the time of writing. Using the same tools as above, that is, word lists and word sketches, we can extract candidate terms and examine their behaviour.

In the English text, *nuclear installation*, *nuclear plant* and *nuclear facility* were used with equal frequency and seem interchangeable. By contrast, according to the patterns in the word sketch for *nucleare* (Table 2.7), we have a clear frontrunner as a potential translation equivalent in Italian, *impianto nucleare*, which is much more frequent than the also acceptable *installazione nucleare*. The Italian corpus is much larger and clearly also covers issues such as nuclear weapons, not the subject of the JRC report. But it is also important to observe that the Italian corpus offers us only one plausible term for *nuclear safety* and *nuclear security*, namely *sicurezza nucleare*, signalling that this term most likely designates both of the concepts referred to in the English text. This hypothesis can be confirmed by searching for both *nuclear safety* and *nuclear security* in other corpora, for example in the Europarl English–Italian parallel corpus (available via Sketch Engine, or from www.statmt.org/europarl/ or via the OPUS collec- tion of corpora at opus.lingfil.uu.se; OPUS, n.d.), where the only Italian translation offered is *sicurezza nucleare*.

Some of the word-sketch patterns for *reattore* (Table 2.8) also reveal terminological correspondences, like *fast reactor* and *reattore veloce*. However, we need to investigate further, looking at the six instances of *raffreddato* (cooling) using the concordancing function to find other compound terms needed. There we see *reattore raffreddato a gas* (gas- cooled reactor), *reattore raffreddate ad acqua* (water-cooled reactor) and *reattore raffreddate ad elio* (helium-cooled reactor). Although not in our data, we might propose *reattore rafreddate a sodio* for *sodium- cooled reactor*, and this term can be verified through other sources.

Finally, as we did for the English ST, we can use Sketch Engine to extract keywords by comparing our *ad-hoc* Italian corpus with a reference corpus for Italian (e.g. the itTenTen corpus, a 3-billion word corpus created by Web crawling and made available in Sketch Engine). As well as generating single-word (simple) term candidates, we can generate lists of keywords based on n-grams to help us to identify com- pound terms. As noted above, bi-grams are sequences of two words, 3-grams are sequences of three words, etc. Since Italian terms are sometimes formed by nouns or noun phrases linked by prepositions, it is worth looking at longer n-grams too, that is, 4-grams or 5-grams. Examples from this corpus data include *ciclo del combustibile* (fuel cycle) and *ciclo del combustibile nucleare* (nuclear fuel cycle).