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 6**

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

**Lecture 6** Creating a corpus from the ST and identifying terms

Having been tasked with the translation, a first stage in our termino- logical research could be to explore the most prevalent terms and their behaviour in the ST. We can do this by creating our own small corpus, consisting of the ST. We can then use some of the software’s functionality to compile a list of single- and multi-word expressions that are likely to be terms. We can also examine the collocational and colligational patterns of their usage.

We can use the ‘Create corpus’ facility in Sketch Engine to upload the ST (a range of file formats are accepted, including PDF). The soft- ware then compiles the corpus, which involves some automatic pro- cessing (parsing, tagging) to produce it in a format on which the corpus analysis tools can be used. The word count tells us there are 21,589 words in the corpus.

The most basic word-list option in any corpus software is to generate a simple **word list**, that is, a list of all the words in the corpus, in order of most frequent to least frequent. Table 2.1 shows the top 20 words from our text (i.e. an extract from the full list that includes all words in the corpus). As is typical for English-language corpora, function words like *the*, *and*, *of*, *in* and *to* are by far the most frequent in our ST. However, the content words that also appear here look like they may form parts of terms designating concepts from our specialized subject domain: *nuclear*, *JRC*, *safety*, *European* and *materials*.

However, for the purposes of identifying terms, this simple word list contains a lot of distracting data, in the form of function words and words repeated in different orthographic or morphological variants (e.g. singular and plural noun forms or different forms of the same verb). We can improve the results of this operation by generating a listing by **lemma** (word stem) rather than by word. Doing so results in word forms such as *is* and *are* being combined into the lemma *be*. Two more content words now appear in the extract of the top 20 items, namely *fuel* and *reactor* – see Table 2.2.

For English and some other languages (Kilgarriff 2013), Sketch Engine has the additional functionality of producing a frequency word list by term, rather than by word or lemma. These are Sketch Engine’s proposals for terms, that is, they are **candidate terms**. As we can see just from looking at the top 20 items from that list, shown in Table 2.3, the function words have disappeared and we have a list made up of items that we would certainly recognize as terms for this domain (e.g. *fuel*, *reactor*, *Euroatom*, *radioactivity*), as well as more general lexis typical of scientific discourse (e.g. *field*, *work*, *support*, *knowledge*). However, we might also note that a frequent candidate term, *JRC*, occurring 254 times in the corpus, is not listed here, so this word list generated by term is a good starting point, but further work is needed.

A refinement on the lemma-based word list, and an option for all languages, is to ask Sketch Engine to extract keywords and possible multiword terms by comparing our nuclear safety corpus with a large reference corpus of English (for this example the reference cor- pus was the enTenTen corpus (2012), which contains over 10 billion words of English obtained as a result of web crawling). This list is present in order of **keyness**. Keyness is a quantitative calculation reflecting the increased frequency of words and phrases in the spe- cialized corpus compared with their frequency in the reference cor- pus (Kilgarriff 2012). All multiword terms in our text, according to Sketch Engine’s analysis of keyness, are listed in Table 2.4. We can identify among them numerous terms that designate specialist nuclear safety concepts. However, we might also spot some phrases which are also prevalent in general language, for example *severe accident*, and some which might be more frequent in scientific discourse but are not specific to nuclear safety, for example *scientific support*, *knowledge management*, *innovation flow*. To illustrate how not all multiword units identified are, in fact, terms, we can see further down the list that *following policy* is proposed as a term.