trans-kom ISSN 1867-4844 http://www.trans-kom.eu
trans-kom ist eine wissenschaftliche Zeitschrift für Translation und Fachkommunikation.

trans-kom 9 [1] (2016): 149-169

Georg Löckinger

Reimagining Terminology Management in an Encyclopaedic Context

Abstract

Recently, language professionals have become increasingly aware that they can make efficient use of tailor-made collections of texts for various activities. This is particularly true for activities involving some kind of text production, such as translation, technical writing or terminology management. After a short introduction (section 1), the present article starts with discussing corpus linguistics, the central concept(s) 'text corpus' as well as its real-world manifestations (section 2). Based on these conceptual fundamentals, we outline two different user profiles with regard to special language reference tools (section 3). Section 4 is about the practical use of text corpora by language professionals. In the following section 5, we take a look beyond the boundaries of text corpora to shed some light on an intelligent combination of languages technologies and language resources. We conclude in section 6 with a brief outlook and a short discussion of research desiderata for the topic at hand.

1 Background

Language professionals working with terminology – such as translators, terminologists, technical writers and interpreters – have a great variety of information sources at their disposal. The most common information sources may be divided into two major groups, according to their purpose:

(1) information sources designed as reference tools for presenting lexical or terminological information: dictionaries, glossaries, nomenclatures, taxonomies, thesauri, etc.

(2) information sources designed for other purposes such as learning and teaching or discourse in expert communities: textbooks, specialist journals, instruction manuals, research reports, etc.

While information sources belonging to the first group usually have properties that facilitate the search for terminological data (designations, definitions, contexts, multimedia content, to name just a few), information sources of the second group typically do not. However, given that they usually describe a more or less specific domain, they also hold valuable information in the terminological sense. This information roughly covers

four levels: object level, concept level, designation level, and context level. Since textbooks, specialist journals or similar information sources have a main purpose other than serving as reference tools, they need to be prepared in a way such that language professionals can use them for terminology management-related activities.

2 Introducing Corpus Linguistics and Text Corpora

While the subject of the present article is practical terminology management, we need to have a closer look at the essential theory-related concepts first. Thus, before discussing user profiles and application scenarios, we will provide a short overview of corpus linguistics and text corpora.

2.1 What Is Corpus Linguistics?

In terminology management, we work with language in different forms (single words, phrases, definitions, or entire texts, etc.). Thus, when it comes to tailor-made collections of texts, we are in the overarching domain of linguistics. Over the past decades, the new subfield corpus linguistics has emerged, which is entirely devoted to real-life texts and their computer-assisted analysis. There are various definitions of 'corpus linguistics' which highlight different characteristics, such as by Biber, Conrad and Reppen (1998: 4) or McEnery and Hardie (2012: 1). Taken together, these definitions clearly suggest that corpus linguistics "is empirical, in that it examines, and draws conclusions from, attested language use, rather than intuitions" (McEnery/Gabrielatos 2006: 34; cf. also Teubert 2007: 41). In other words, corpus linguistics serves both to discover new features of natural language that might not be evident from introspection and to verify or falsify assumptions about natural language by closely observing its use (cf. Teubert/ Čermáková 2007: 137).[1] A body of text that is examined using methods and tools of corpus linguistics is usually called "text corpus" or "corpus".

2.2 What Is a Text Corpus?

As a rule, the object of study in corpus linguistics is something called "text corpus" or "corpus". While the two terms are often used as synonyms, we do not use the shorter one "corpus" in the present article: for our purposes, we are mainly concerned with content consisting of written characters, as opposed to non-textual material such as audio or video files. In this case, "text corpus" is the more precise and transparent term in the terminological sense.

In the literature on terminology management and corpus linguistics, we can find a wide range of definitions of the concept 'text corpus'. Some of them are narrower in that they put the emphasis on the scientific nature of linguistics or on language technology aspects:

---

[1] While language use is studied in various other subfields of linguistics too, corpus linguistics has an explicit focus on empirical methods.

- "collection[s] [of linguistic data] constructed according to various linguistic criteria such as representativeness and balance across a given domain, set of languages, etc." (Expert Advisory Group 2000)
- "systematic collection of machine-readable texts […] or parts of text prepared, coded and stored according to predefined rules" (ISO 1087-2 2000: 2)
- "collection of more than one text […] [with] more specific connotations … [which] may be considered under four main headings:
  - sampling and representativeness
  - finite size
  - machine-readable form
  - a standard reference" (McEnery/Wilson 2001: 29)
- "finite collection of machine-readable text, sampled to be maximally representative of a language or variety" (McEnery/Wilson 2001: 197)
- "collection of texts or parts of texts upon which some general linguistic analysis can be conducted" (Meyer 2002: xi)
- "collection of naturally occurring language texts in electronic form, often compiled according to specific design criteria and typically containing many million of words" (Teubert 2007: 140)

On the other hand, some definitions represent broader concepts, which are more appropriate in our context since we are discussing the practical activity called terminology management:

- "large and principled collection of natural texts" (Biber/Conrad/Reppen 1998: 4)
- "collection of language data brought together for analysis" (ISO 1087-1 2000: 11)
- "collection of linguistic data, including written, spoken, or both, in one or multiple languages" (Expert Advisory Group 2000)
- "body of machine-readable text" (McEnery/Wilson 2001: 197)

In the above more general definitions, we see an overarching common characteristic <u>collection of language data or text[s]</u>,[2] which means that text corpora are usually more than a single sentence or paragraph, for instance. Also, it is obvious that the act of collecting is not and end in itself, but serves to facilitate some sort of subsequent <u>analysis</u>. In terminology management, this may include tasks such as comparing definitions, finding contexts, examining the frequencies of designations, etc. Finally, what we need to analyse is available electronically today in most cases: the relevant content is <u>machine-readable</u>, i.e. we can access and process it using computers. In the following, when discussing text corpora, we assume a concept 'text corpus' that consists of at least the above three characteristics.

---

[2] Underline formatting in accordance with relevant national and international standards such as ÖNORM A 2704 (2015) and ISO 704 (2009).

2.3 What Types of Text Corpora Do Exist?

Based on the concept 'text corpus' as described in the previous section, information sources which are not primarily designed as reference tools (see section 1, point 2) are packed into text corpora of various types, including the following:

- monolingual, bilingual or multilingual
- parallel or comparable (cf., for instance, Teubert/Čermáková 2007: 73ff; Aijmer 2008)
- written or spoken (cf., for instance, Granger 2008: 262-263; Hundt 2008: 169-170)
- domain-specific or cross-domain
- pre-defined or tailor-made
- non-web or web (cf., for instance, Teubert/Čermáková 2007: 76-77)
- raw or pre-processed
- static or dynamic
- freely available or paid

In the following, we describe in greater detail some types of text corpora that might be less known by, yet helpful for language professionals.

2.3.1 Parallel vs. Comparable

The terms "parallel corpus" and "comparable corpus" are not as transparent in the terminological sense as one might wish. Also, various publications use them for denoting overlapping or even opposing concepts. In the present article, we use the differentiation used by Aijmer. According to her, both parallel and comparable corpora include content in at least two languages. While a parallel corpus contains human-authored texts in a source language and their equivalent human-translated texts in a target language, a comparable corpus consists of texts in different languages that are interrelated with regard to the domain, text type, date of creation, etc. For language professionals, it depends on their specific needs (see section 4) whether they need to use a parallel or comparable corpus, or a combination of both (Aijmer 2008).

Example

The *European Centre for Disease Prevention and Control Translation Memory* (*ECDC-TM*) is a parallel corpus. This multilingual text corpus for the health domain holds human-authored and human-translated content in 25 languages; it is pre-defined (its content has been selected by the issuing organisation), it is pre-processed (segmented and fed into Translation Memory eXchange files) and is freely available at the website of the European Commission's (ca. 2014) Joint Research Centre. *ECDC-TM* may be particularly useful for translators and technical writers working with health-related texts in one or several of the languages covered by *ECDC-TM*.

2.3.2 Pre-defined vs. Tailor-made

This dichotomy refers to the question whether a corpus has been compiled with a given target audience in mind from the outset or whether language professionals themselves create a corpus for their own specific purposes. While pre-defined corpora are often ready to use, even free via the Internet, tailor-made corpora need to be compiled taking into account the relevant conditions and parameters (see section 4.1). Texts that lend themselves to corpus exploitation are often not available publicly, or they are internal and the only source of information anyway. In such situations, user-specific compilation and maintenance is necessary.

Example

The *Karl-May-Korpus* provided by the European Language Resources Association (ELRA) is a pre-defined corpus. It has been created for people who need to examine or use the language used by the German author Karl May. It is monolingual (German), pre-processed (SGML-tagged) and available for a fee, depending on the type of user you are; for more detailed information, see the ELRA website (European Language Resources Association 2008). The *Karl-May-Korpus* may be interesting for literary translators working with texts by Karl May or for linguists examining the language used by the German author.

2.3.3 Raw vs. Pre-processed

A text corpus may be nothing more than several flat files in their original form, such as three PDF files grouped into one file folder. In that case, a text corpus is raw, i.e. the texts concerned have not been treated by any natural language processing method. For sophisticated analyses, it is paramount that a text corpus undergoes some pre-processing (or an entire pre-processing pipeline) before it is exploited by users. For example, let us assume a technical writer needs to know in what and how many adjective/noun combinations the noun “coffeemaker” is used in a domain-specific text corpus on coffee preparation. To be of maximum use to the technical writer, the text corpus needs to be treated by means of part-of-speech tagging, i.e. a computer application annotates (“tags”) every individual word in the corpus according to its word class.

3 Creating Tailor-Made Resources: Two Selected User Profiles

As discussed in the previous section, various types of text corpora may be used in language professionals’ daily work. However, these alone cannot usually cover the wide range of language professionals’ information needs. To determine what information should be included in special language reference tools for the relevant (types of) users, it is necessary to create a profile of them stating what activity they carry out, what competences should be considered, etc. The discussion below is a general one and thus

applicable to many specific application scenarios, independently of any particular domain, language or language combination.

Löckinger (2014: 63ff) developed a user profile for translators based on the existing literature on user modelling for reference tools, in particular Geeb (1998: 40ff). In a refined form and displayed using the concept modelling conventions according to ISO 24156-1 (2014), this user profile looks like follows (figure 1):

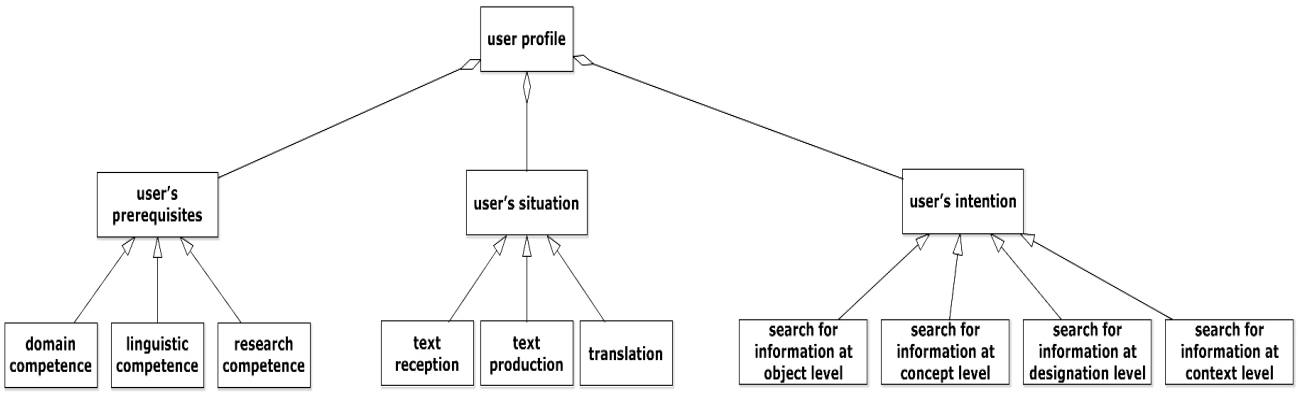

Fig. 1: Profile of translators as users of special language reference tools (translated and refined version of Löckinger 2014: 63).[3]

For technical writers, the user profile is very similar, as described in Löckinger (2015: 12ff) (figure 2):

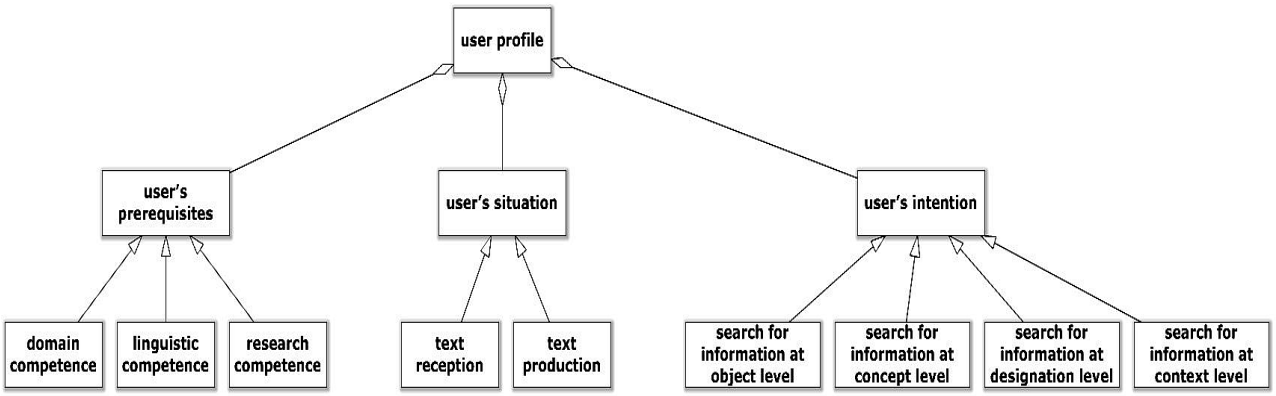

Fig. 2: Profile of technical writers, as presented in Löckinger (2015: 13).[4]

---

3 Since the user profile is displayed in the form of a concept model according to ISO 24156-1 (2014), the conventions of the Unified Modeling Language are used: the symbol ◇—— represents a partitive relation in the terminological sense. For instance, 'user's intention' is one of three parts of 'user profile'. The symbol ◁—— represents a generic relation in the terminological sense. Thus, 'search for information at designation level' is a kind of 'user's intention', etc.

4 See footnote 2.

As discussed in Löckinger, we can assume that there are only a few general differences between the user profiles 'translator' and 'technical writer':

(1) Given the additional user's situation 'translation', reference tools for translators are often bilingual or multilingual, while monolingual language resources might be sufficient for technical writers.

(2) Some types of data, i.e. data categories, are specific to translators, for instance information on the degree of correspondence between concepts or the degree of equivalence between designations. (Löckinger 2015: 15ff)

By contrast, especially with regard to the user's intention, translators and technical writers generally have very similar needs (always depending on their individual education/training and professional experience, of course): they might search for information at the object level, such as details on how to use a rotary drill; information at the concept level, such as a definition or similar description of the concept 'rotary drill'; information at the designation level, such as about the frequency of use of the term "rotary drill" in corporate texts; and/or information at the context level, such as wording patterns used to describe the safe use of rotary drills.

In an empirical study with professional translators, Löckinger compared the different rates of success – success meaning reference tool users finding the information needed – for two different translation-oriented language resources covering the same domain (security policy). At one stage, the participants had to translate a text from English into German using a domain-specific terminological database in these languages, which included bilingual concept diagrams. At the next stage, the participants could additionally make use of tailor-made text corpora consisting of authentic English and German texts, i.e. non-translated material. It turned out that for the six translators participating, the rate of success increased, partly considerably, with regard to all information categories (such as designations, grammar and spelling information, contexts, definitions and similar concept descriptions, etc.) (Löckinger 2014: 151ff).

Another important finding was that the participants generally appreciated the text corpora available within the same user interface as other terminological and lexical data. While due to the research design the above-mentioned results cannot be extrapolated to all translators, let alone all language professionals, these findings support the idea that text corpora might be helpful aids for translators and technical writers (cf. Löckinger 2014: 200ff).

For reference tools created and designed for these target groups, it should however be borne in mind that more is needed than mere text corpora. Based on the study above and on real-world experience, a combined resource ("dynamic […] terminology and full-text database", Löckinger 2014: 316) will probably deliver better results.

4 Using Text Corpora in Practice

Having discussed why it is paramount to develop (or make use of) user profiles before creating reference tools for language professionals, we can now have a closer look at the way in which text corpora should be compiled or selected in practice. Whether a text corpus must be created from scratch or can be readily used, those responsible for taking decisions (language professionals themselves or personnel supporting them) should take into account both the relevant general user profile (see section 3) and the specific conditions (external factors) and parameters (text corpus properties) outlined below. What follows can thus serve as a set of practical guidelines on the compilation and selection of text corpora. In addition to that, we also discuss potential application scenarios for translators and technical writers, illustrating our point with appropriate examples.

4.1 Conditions and Parameters for Selecting and Compiling Text Corpora

When selecting or compiling a text corpus, language professionals should consider the following *conditions (external factors)*:

- *(type of) user*, e.g. technical writers, professional translators, terminologists, or experts consulting a terminological database; see also section 3 on user profiles
- *(type of) information deficit to be covered*, e.g. target-language equivalents of a source-language term, ad-hoc information deficit regarding encyclopaedic information at the object level, or a long-term information deficit at concept and designation levels experienced by many people within an organisation
- *purpose of text corpus use*, e.g. quality assurance, exploitation for statistical purposes, translation-oriented terminology management, or monolingual text production
- *time and resources available*, e.g. only one person for a very short timeframe, or a terminology management team responsible for an entire organisation
- *existing language technology processes*, e.g. automatic translation memory retrieval, pre-processing or post-processing of documents with part-of-speech tagging or similar annotation techniques
- *existing language technology tools*, e.g. no tools in use, or off-the-shelf terminology management system used across the board

Depending on the actual needs of a terminology management project and terminology users, a text corpus should then be compiled or selected based on detailed parameters (properties of text corpora) reflecting the conditions enumerated above.

What follows is a non-exhaustive list:

- language coverage
- domain coverage
- content type
- customization
- medium
- pre-processing
- dynamism
- costs
- …

Example

A professional translator is commissioned to translate an instruction manual from English into German. The instruction manual is about a new coffeemaker by a company that has been selling several coffeemakers of the same series for years. Since the translator is not yet very familiar with the client's corporate language in German texts, he/she decides to combine all existing instruction manuals in the target language German in one raw text corpus. These texts will typically be available for free at the company's website or be made available to the translator upon his/her request. By means of this text corpus of instruction manuals, the translator can examine how existing German texts about the company's coffeemaker series are worded, what terms and names are used where and how often, etc. Using a dedicated tool, the translator will find the necessary information faster than using standard software (e.g. text processor, PDF reader or similar): thus, the use of text corpus management or concordance software is highly recommended in such a case; see section 4.2 for examples.

For the example above, determining the specific conditions (external factors) results in the following list:

- *type of user:* professional translator (see the translator user profile in section 3)
- *type of information deficit:* designation, concept, and context levels
- *purpose of text corpus use:* monolingual text production within a translation setting
- *time/resources available:* one person (translator himself/herself)
- *existing language technology processes:* unknown
- *existing language technology tools:* unknown

Taking these condition values, we can derive *parameter* values that reflect in detail what properties the text corpus concerned must have:

- language coverage: monolingual (German)
- domain coverage: coffee preparation (specific to the company in question)
- content type: parts of a parallel corpus

- customisation: tailor-made by the translator himself/herself
- medium: non-web
- pre-processing: unknown
- dynamism: dynamic resource (translator can add content to and remove content from the text corpus)
- costs: no immediate costs for the resource itself

To sum up, the selection or compilation of a text corpus roughly follows a four-step approach:

(1) Using an existing user profile or creating a new one.

(2) Determining the conditions (external factors) under which a text corpus is to be used.

(3) Reflecting these conditions in parameter values which correspond to the properties of the text corpus needed.

(4) Creating or using an appropriate text corpus.

4.2 Possible Application Scenarios

Language professionals may use text corpora for a wide variety of activities such as those listed below. For illustration purposes, screenshots are provided to show how an activity may look like in designated text corpus management tools or text corpus portals.

Searching for equivalents in one language of a designation in another language (cf., for instance, Nuopponen 1996: 1070)

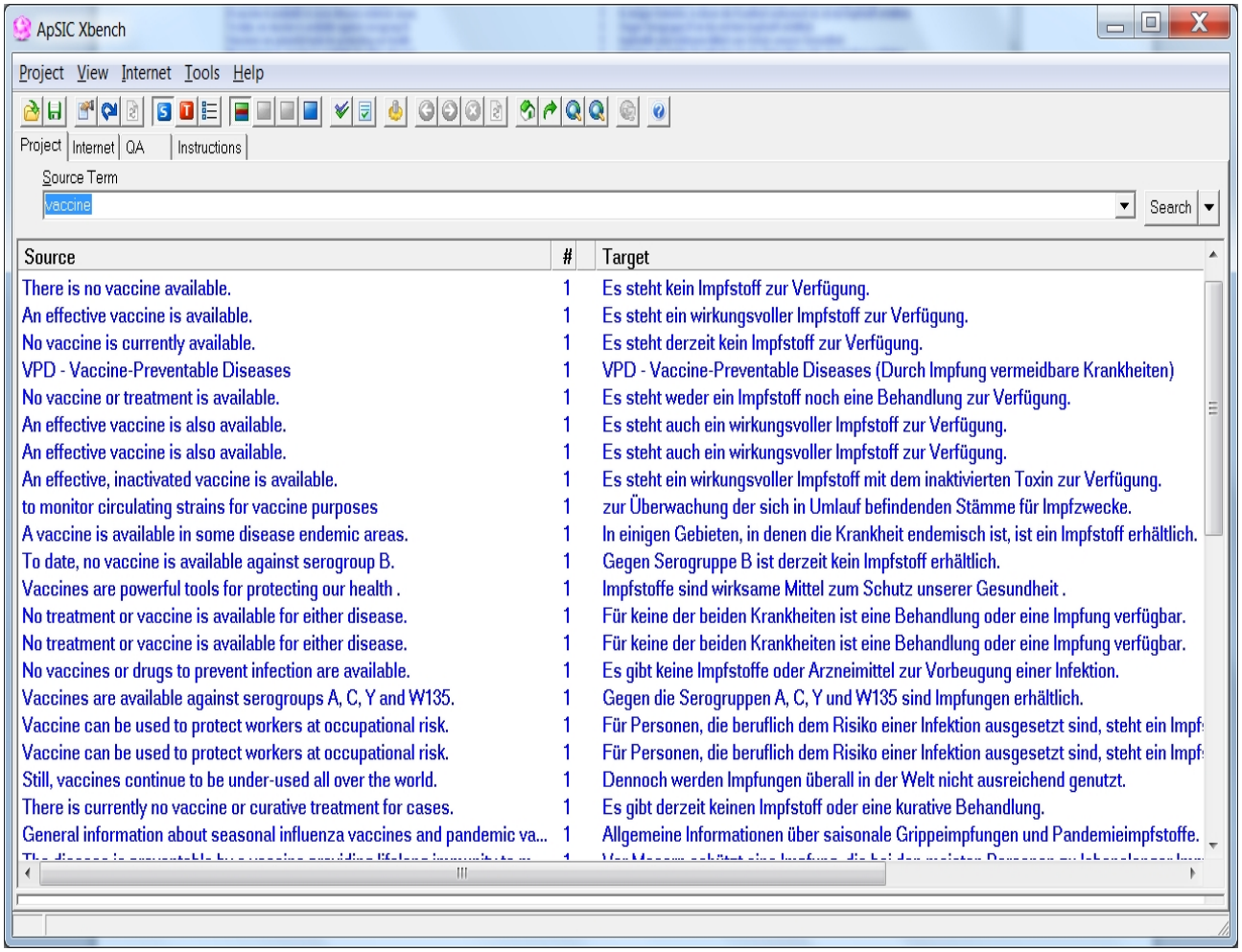

Fig. 3: Searching for German equivalents of the English term “vaccine” in parts of the ECDC-TM parallel corpus (European Commission ca. 2014), using the tool *ApSIC Xbench 2.9*.

Carrying out quality assurance within monolingual or multilingual text production processes, e.g. crosschecking texts for any terminological inconsistencies or mistakes

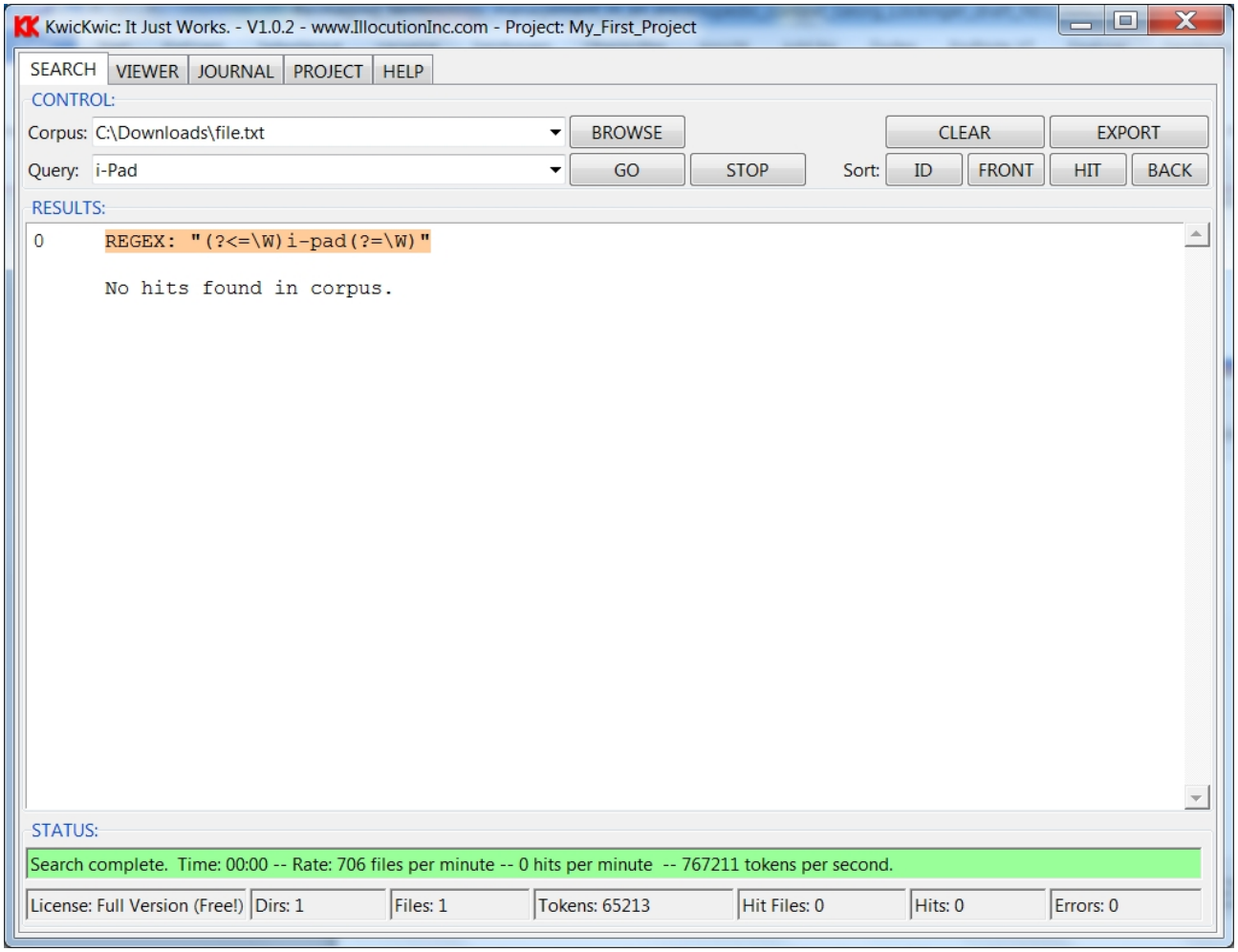

Fig. 4: Searching for unwanted hyphenated forms of a proper name, using the tool *KwicKwic 1.0.2*.

trans-kom 9 [1] (2016): 149-169

Exploring the usage of designations in real-life texts (cf., for instance, Leech/ Nesi 1999: 303)

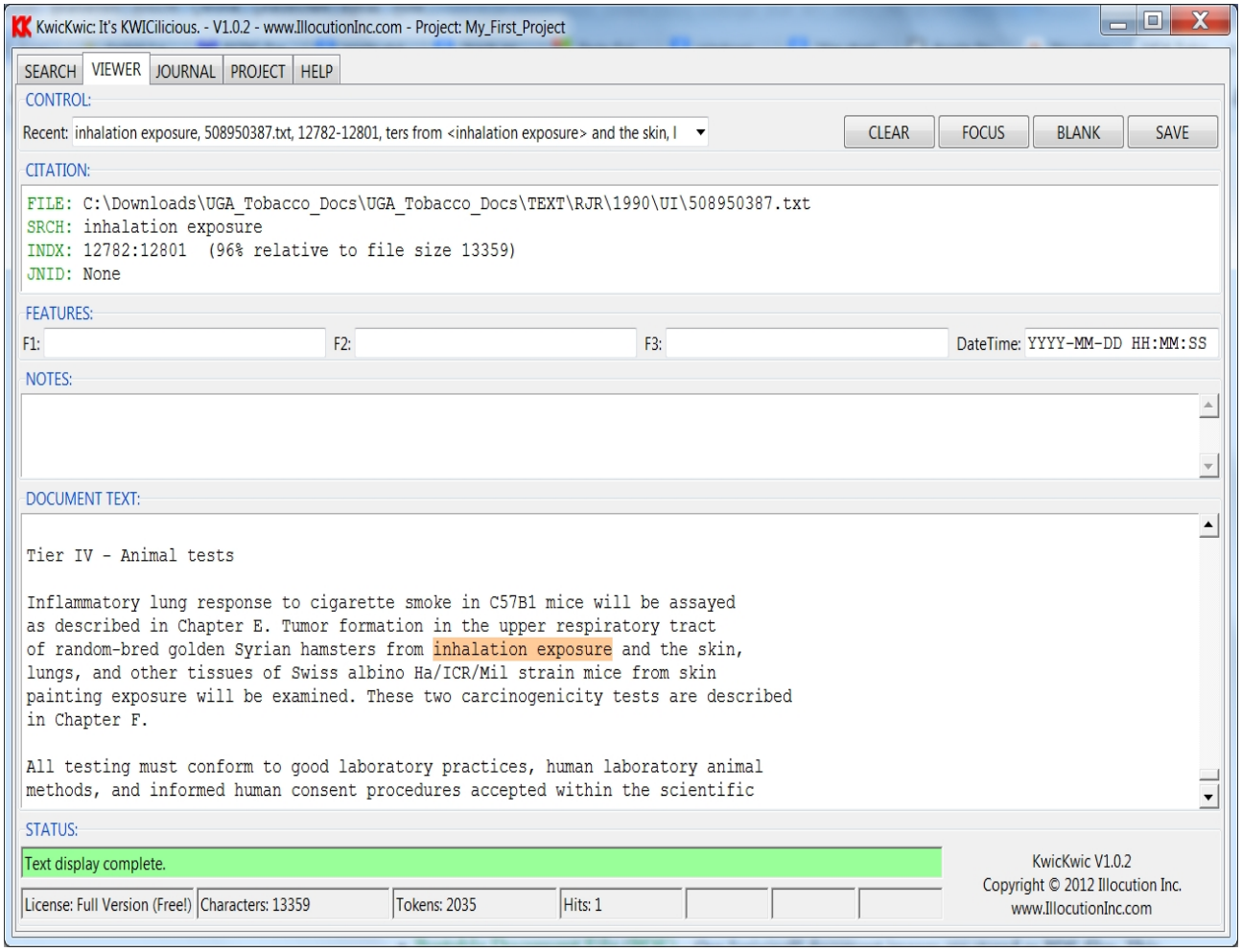

Fig. 5: Investigating the usage of the term “inhalation exposure” in the University of Georgia (n.d.) Tobacco Documents Corpus, using the tool *KwicKwic 1.0.2*.

Examining texts for the frequency of use of designations (cf., for instance, Biber/Conrad/Reppen 1998: 28ff), or comparing the usage of two or more designations

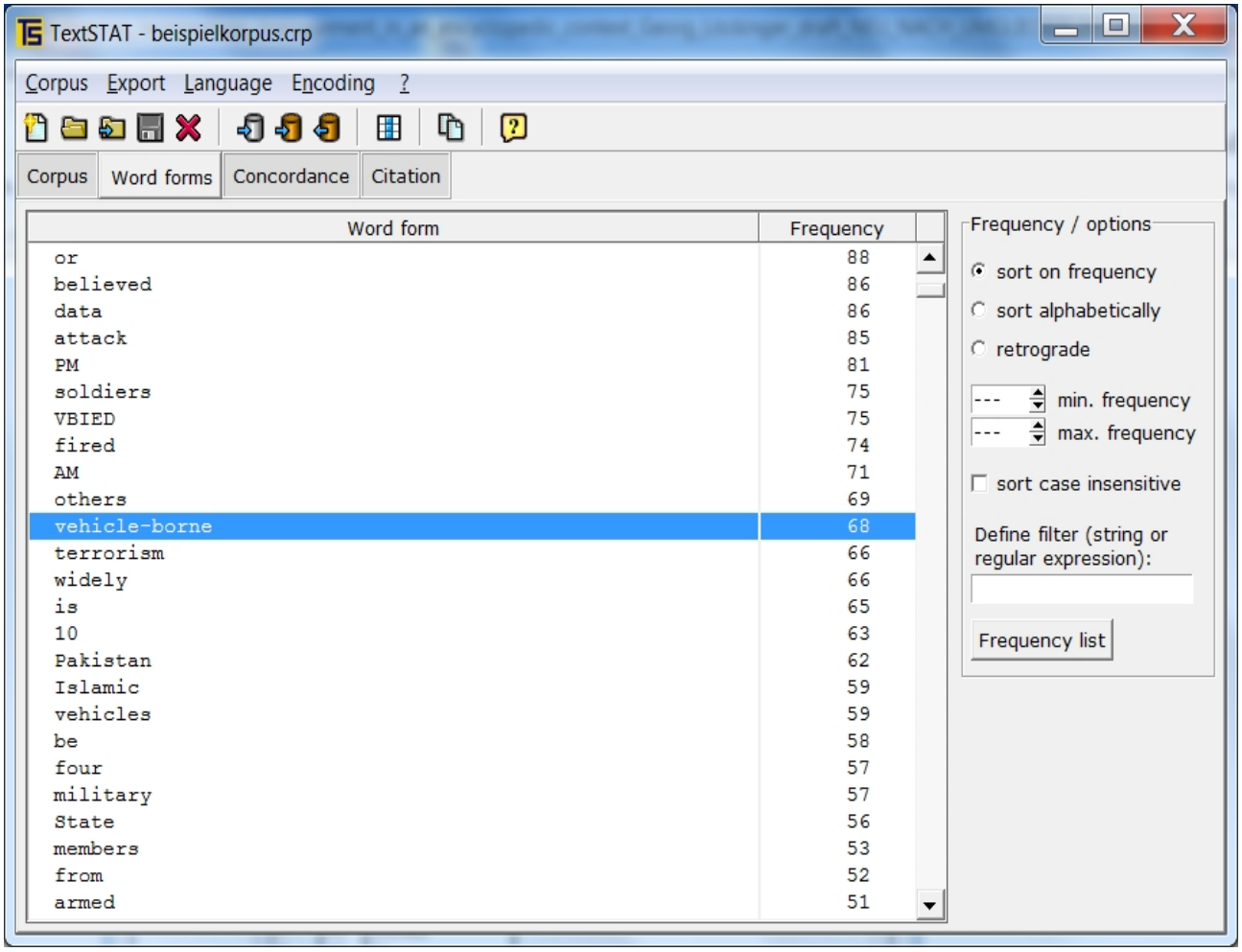

Fig. 6: Checking the frequency of use of “vehicle-borne” in National Counterterrorism Center (2009), using the tool *TextSTAT 2.9c*.

Examining the stylistic and syntactic behaviour of designations (cf., for instance, McEnery/Wilson 2001: 108)

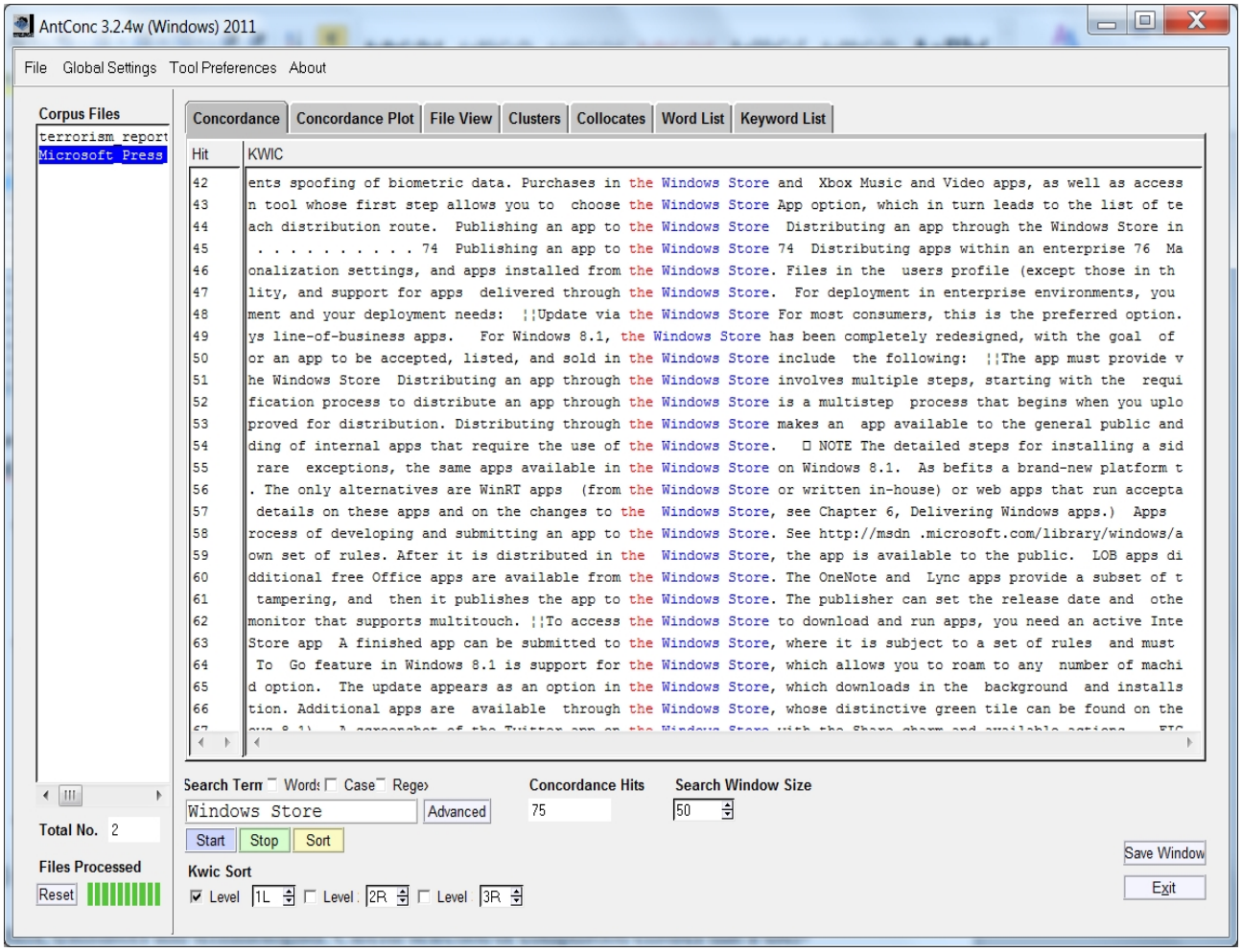

Fig. 7: Looking at the syntactic behaviour of a proper name in Bott (2013), using the tool *AntConc 3.2.4w*.

Exploring domain-specific or text-type conventions typical for the texts in question (cf. Biber/Conrad/Reppen 1998: 45ff)

CORPUS OF CONTEMPORARY AMERICAN ENGLISH — G LÖCKINGER

450 MILLION WORDS, 1990-2012 [DOWNLOAD ALL 190,000 TEXTS] -- START -- ? history | lists | logout

1	DIRECT NEGOTIATIONS	32
2	INTENSE NEGOTIATIONS	29
3	SERIOUS NEGOTIATIONS	23

KEYWORD IN CONTEXT DISPLAY — Help / information / contact

SECTION: NEWSPAPER (32)

CLICK FOR MORE CONTEXT [?] SAVE LIST CHOOSE LIST --------- CREATE NEW LIST [?]

1	2012	NEWS	WashPost	A	B	C	ahead, and administration officials have said it is intended as a venue for **direct negotiations** between the Afghan government and the insurgents. # This week, Marc Grossman,
2	2011	NEWS	AssocPress	A	B	C	process. # In a statement, Netanyahu called on Abbas to begin " **direct negotiations** in New York and continue them in Jerusalem and Ramallah. " But the statement
3	2011	NEWS	AssocPress	A	B	C	" Juppe said. " The only way to settle the Israeli-Palestinian problem is **direct negotiations**. " # Venezuelan President Hugo Chavez, meanwhile, told Secretary-General Ban Ki-moon i
4	2011	NEWS	AssocPress	A	B	C	official recognition, the administration has been working intensively behind the scenes to restart **direct negotiations** between Israelis and Palestinians and to persuade Palestinian lea
5	2011	NEWS	CSMonitor	A	B	C	" within an month " at a preparatory meeting to set an agenda for **direct negotiations**. The Quartet said it would expect talks to have several interim goals but to
6	2011	NEWS	CSMonitor	A	B	C	US is " very pleased " the Quartet had reached an agreement for restarting **direct negotiations** " without delay or preconditions. " # Leaving unaddressed the Palestinian statehood a
7	2011	NEWS	CSMonitor	A	B	C	administration's view that " a just and lasting peace can only come through **direct negotiations**. " # Also addressing the press, Quartet representative Tony Blair said he thought
8	2011	NEWS	CSMonitor	A	B	C	representative Tony Blair said he thought the Palestinian bid in the Security Council and **direct negotiations** could be " combined " and proceed concurrently. # Abbas insisted in his
9	2011	NEWS	CSMonitor	A	B	C	his speech that the statehood bid is not closing the door on resumption of **direct negotiations** with Israel, but he repeated the conditions for returning to talks that Israeli Prime
10	2011	NEWS	CSMonitor	A	B	C	the West Bank? # While Mr. Oren and other Israeli officials insist on **direct negotiations** as the only path to peace, Palestinian leaders are exasperated with nearly two decades
11	2010	NEWS	NYTimes	A	B	C	' in a reference to the West Bank and East Jerusalem. # The **direct negotiations**, scheduled to start Thursday, will be the first in 20 months. #
12	2008	NEWS	CSMonitor	A	B	C	Kabul may be hoping to use these talks as a starting point for future **direct negotiations** with senior leadership. # Some who attended the Mecca meetings say that future meetings
13	2006	NEWS	USAToday	A	B	C	simply blackmail, Bob. Kim Jong Il wants us to again engage in **direct negotiations** with him as a ploy for winning aid, but his past promises have been
14	2005	NEWS	Atlanta	A	B	C	. But the unilateral Gaza withdrawal was basically a military redeployment designed to preclude **direct negotiations** with the Palestinians over land they claim for a future state. # " I
15	2004	NEWS	NYTimes	A	B	C	continue that work. Mr. Doctoroff said that he has not been in any **direct negotiations** with Mr. Ross or Related and that he is not a member of the group
16	2004	NEWS	Atlanta	A	B	C	sides back to the negotiating table, but EU regulators still balk at resuming **direct negotiations**, officials said Thursday. The European Court of First Instance is expected to decide
17	2003	NEWS	WashPost	A	B	C	billion fund to encourage private health plans and prohibit federal Medicare officials from conducting **direct negotiations** over prices with pharmaceutical companies. # " This isn't the
18	2000	NEWS	WashPost	A	B	C	a quarter of its water. # Events in the Israeli-Palestinian peace process since **direct negotiations** began in 1993: # Land # By the end of 1995, Israeli forces
19	1998	NEWS	CSMonitor	A	B	C	to withdraw from 13.1 percent more of the West Bank. # The first **direct negotiations** in four years between North and South Korea appeared to collapse in Beijing, with
20	1997	NEWS	CSMonitor	A	B	C	fix. " He said the US can not advance the peace process without **direct negotiations** between them. Responding, Palestinian Authority President Arafat complained that the US had fa
21	1994	NEWS	WashPost	A	B	C	said, however, that " our view has always been that we support **direct negotiations** between the two sides. If it's true, we would welcome that summit
22	1992	NEWS	CSMonitor	A	B	C	warring partners before it is too late. There are no prospects now for **direct negotiations**, says Zori Balayan, a member of the former Soviet parliament from Karabakh and
23	1991	NEWS	NYTimes	A	B	C	is a moment of historic opportunity. For 43 years, Israel has sought **direct negotiations** with its neighbors, and it has been right to do so. # "
24	1991	NEWS	NYTimes	A	B	C	its neighbors, and it has been right to do so. # " **Direct negotiations** are the only way to solve problems and the only way to secure peace,
25	1991	NEWS	NYTimes	A	B	C	that the conference could occasionally reconvene after Israel and the Arab countries had begun **direct negotiations**. # The Israelis have opposed reconvening the full peace conferen
26	1991	NEWS	NYTimes	A	B	C	time that we have an opportunity to negotiate as we have wanted to -- **direct negotiations**, without a mediator, face-to-face, without preconditions -- with our Arab neighbors.
27	1991	NEWS	NYTimes	A	B	C	real peace, " but the Syrians rejected that wording and notion. These **direct negotiations**, the invitation says, should follow " immediately " after the last day of
28	1991	NEWS	AssocPress	A	B	C	for peace. We support the concept of a peace conference starting parallel and **direct negotiations** between Israel and representative Palestinians on the one hand and Israel and the

Fig. 8: Examining collocations including the noun “negotiation” in newspapers, based on the part-of-speech criterion ‘preceding adjective’, using the Brigham Young University (n.d.) corpus portal (Corpus of Contemporary American English).

Overall, text corpora may be very helpful for language professionals such as translators and technical writers. Careful selection or compilation ensures that a user-specific language resource is at their disposal for terminology research and similar language-related activities. The above application scenarios provide a glimpse of the potential that corpora hold for language professionals’ daily work.

trans-kom 9 [1] (2016): 149-169

5 Going Beyond Text Corpora: Language Technologies and Resources across the Board

As indicated in section 3, language professionals need special language reference tools that contain more than mere text corpora. For instance, translation memory repositories or terminological databases are often among the resources that language professionals consult when searching for terminological or lexical information. What lies behind the reference tools themselves is a more or less diverse range of language technologies, many of which are usually less known and used by language professionals than by computer linguists or researchers working with language technologies.

Thus, the following term cloud for language technologies (figure 9) is intended to give a brief overview that might be helpful for language professionals. It is essential that providers of special language reference tools choose wisely according to the needs of the relevant users (see section 3 on user profiles) and deliver an entire technology pipeline guiding users through their language-related activities.[5]

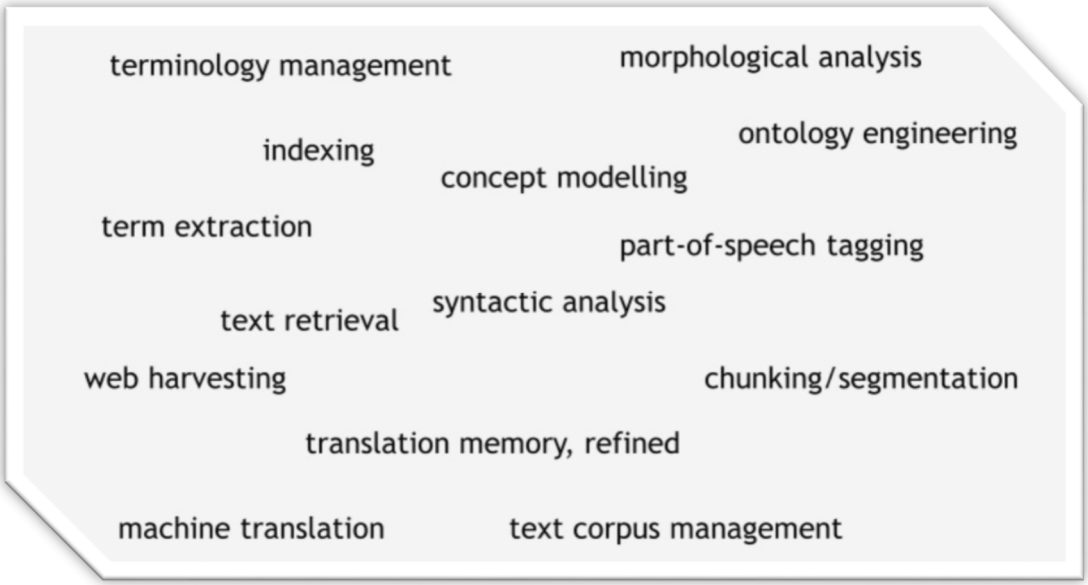

Fig. 9: Term cloud “language technologies for language professionals”.

With regard to language technologies, we can state that there are also several dichotomies to consider when it comes to the user-oriented provision of language resources (figure 10):

[5] For instance, ÖNORM A 2704 (2015) and ISO 704 (2009) describe the main activities and tasks involved in terminology management, which may be used as a basis for developing new or adapting existing tools.

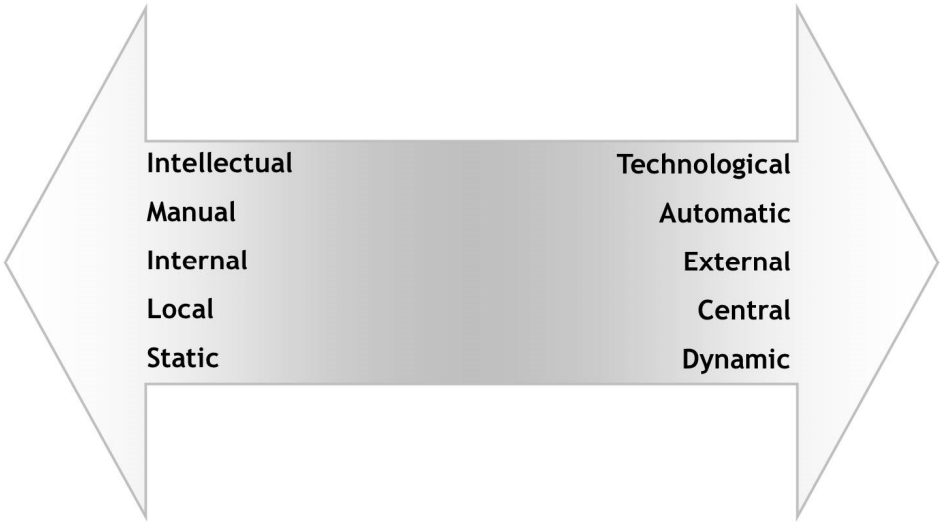

Fig. 10: Dichotomies to consider when providing language resources and language technologies.

- *Intellectual/Technological*. It is paramount to let the machine do what it can do better than the users, and vice versa. It does not make any sense for a human trying to count all frequencies of all words in a text corpus consisting of a thousand documents with hundreds of pages each. Nor would it be realistic at this point to expect software to automatically compare several definitions of possibly overlapping concepts.
- *Manual/Automatic*. Small numbers of texts can often be collected much faster in manual mode (downloading files from the Internet, putting them into a dedicated folder structure) than in an automatic mode that one would have to program first.
- *Internal/External*. Third, the content needed for a specific text corpus might be scattered over different places that are both internal and external to an organisation. From a slightly different perspective, this dichotomy also refers to the question whether language resources should be kept in internal and/or external data repositories.
- *Local/Central*. Using language resources and language technologies poses the question how they should be handled from an IT management point of view. Local ownership provides the individual users with more flexibility for the details of their activities, while as a rule central service provision and storage are more robust and deliver a better performance, in particular with regard to large amounts of data and CPU-intensive queries.
- *Static/Dynamic*. Finally, it depends on the users and their needs whether the language resources provided are static, i.e. created once and not modified thereafter, or rather dynamic, i.e. they are added to, reduced and/or changed at regular intervals or from time to time, as appropriate.

6 Conclusion and Outlook

In the past years and decades, text corpora of different types and quality have emerged, and many free text corpora are even easily accessible via the Internet. However, text corpora are still the least well recognized and researched type of reference tool for language professionals. In addition to that, they are substantially underestimated with regard to their potential use for people working with language in general. So, text corpora should be exploited more systematically and should be part and parcel of language professionals' working environments.

When it comes to the integration of text corpora, many traditional tools for language resource management, such as translation memory systems, do not yet exploit text corpora as systematically as they could. However, there are some language technology companies whose tools do enable users to combine the main resource types (terminological data, translation memory data and text corpora) within one and the same user interface, which is essential from a usability point of view.[6]

Text corpora still hold a number of research desiderata, for instance language professionals' awareness of text corpora and text corpus management technologies or language professionals' use of text corpora in their daily work. Furthermore, a detailed market overview of free and commercial tools, which is continuously maintained, would be valuable for researchers and language professionals alike.

Future research on these issues, in particular empirical studies, might further explore the topic at hand. Ideally, this will result in a more detailed and more comprehensive picture of terminology management in a reimagined, encyclopaedic context.

Acknowledgment

The author is very grateful to Sue Ellen Wright, who provided the original idea for this article.

References

Aijmer, Karin (2008): "Parallel and Comparable Corpora." Anke Lüdeling, Merja Kytö (eds): *Corpus Linguistics. An International Handbook.* Vol. 1. Berlin: Walter de Gruyter, 275-292

Biber, Douglas; Susan Conrad; Randi Reppen (1998): *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press

Bott, Ed (2013): *Introducing Windows 8.1 for IT Professionals. Technical Overview*. Redmond: Microsoft Press – http://download.microsoft.com/download/E/5/3/E5395265-D2CD-4451-A2BB-B4504C000E80/Microsoft_Press_ebook_Introducing_Windows_ITPro_PDF.pdf (11 April 2016)

Brigham Young University (n.d.): corpus.byu.edu – http://corpus.byu.edu/ (06 June 2016)

[6] A recent example is Kilgray's memoQ, a tool for translators, which contains a text corpus feature called LiveDocs.

European Commission (ca. 2014): Joint Research Centre: ECDC-Translation Memory – https://ec.europa.eu/jrc/en/language-technologies/ecdc-translation-memory (06.06.2016)

European Language Resources Association (2008): Karl-May-Korpus (KMK) – http://catalog.elra.info/product_info.php?products_id=450 (06 June 2016)

Expert Advisory Group on Language Engineering Standards (2000): *Corpus Encoding Standard* – http://www.cs.vassar.edu/CES/CES1.html (11 April 2016)

Geeb, Franziskus (1998): *Semantische und enzyklopädische Informationen in Fachwörterbüchern. Eine Untersuchung zu fachinformativen Informationstypen mit besonderer Berücksichtigung wortgebundener Darstellungsformen*. Århus: Århus School of Business

Granger, Sylviane (2008): "Learner corpora." Anke Lüdeling, Merja Kytö (eds): *Corpus Linguistics. An International Handbook.* Vol. 1. Berlin: Walter de Gruyter, 259-275

Hundt, Marianne (2008): "Text corpora." Anke Lüdeling, Merja Kytö (eds): *Corpus Linguistics. An International Handbook*. Vol. 1. Berlin: Walter de Gruyter, 168-187

ISO 1087-1 (2000). *Terminology Work – Vocabulary.* Part 1: *Theory and Application*. Geneva: International Organization for Standardization

ISO 1087-2 (2000): *Terminology Work – Vocabulary.* Part 2: *Computer Applications*. Geneva: International Organization for Standardization

ISO 24156-1 (2014): *Graphic Notations for Concept Modelling in Terminology Work and its Relationship with UML – Part 1: Guidelines for Using UML Notation in Terminology Work*. Geneva: International Organization for Standardization

Leech, Geoffrey; Hilary Nesi (1999): "Moving towards Perfection: The Learner's (Electronic) Dictionary of the Future." Thomas Herbst, Kerstin Popp (eds): *The Perfect Learners' Dictionary (?)*. Tübingen: Narr, 295-306

Löckinger, Georg (2014). *Übersetzungsorientierte Fachwörterbücher. Entwicklung und Erprobung eines innovativen Modells*. Berlin: Frank & Timme – http://research.fh-ooe.at/de/publication/4233 (11 April 2016)

trans-kom **ISSN 1867-4844**

trans-kom ist eine wissenschaftliche Zeitschrift für Translation und Fachkommunikation.

trans-kom veröffentlicht Forschungsergebnisse und wissenschaftliche Diskussionsbeiträge zu Themen des Übersetzens und Dolmetschens, der Fachkommunikation, der Technikkommunikation, der Fachsprachen, der Terminologie und verwandter Gebiete.

Beiträge können in deutscher, englischer, französischer oder spanischer Sprache eingereicht werden. Sie müssen nach den Publikationsrichtlinien der Zeitschrift gestaltet sein. Diese Richtlinien können von der **trans-kom**-Website heruntergeladen werden. Alle Beiträge werden vor der Veröffentlichung anonym begutachtet.

trans-kom wird ausschließlich im Internet publiziert: http://www.trans-kom.eu

Redaktion

Leona Van Vaerenbergh
University of Antwerp
Arts and Philosophy
Applied Linguistics / Translation and Interpreting
Schilderstraat 41
B-2000 Antwerpen
Belgien
Leona.VanVaerenbergh@uantwerpen.be

Klaus Schubert
Universität Hildesheim
Institut für Übersetzungswissenschaft
und Fachkommunikation
Universitätsplatz 1
D-31141 Hildesheim
Deutschland
klaus.schubert@uni-hildesheim.de

Löckinger, Georg (2015): "Designing State-of-the-art Reference Tools for Technical Writers." European Association for Technical Communication – tekom Europe e.V. (ed.): *Proceedings of the European Academic Colloquium on Technical Communication 2014*. Stuttgart: tcworld GmbH, 11-24

McEnery, Tony; Costas Gabrielatos (2006): "English Corpus Linguistics." Bas Aarts, April McMahon (eds): *The Handbook of English Linguistics*. Malden: Blackwell, 33-71

McEnery, Tony; Andrew Hardie (2012): *Corpus Linguistics. Method, Theory and Practice*. Cambridge: Cambridge University Press

McEnery, Tony; Andrew Wilson (2001): *Corpus Linguistics. An Introduction*. 2nd ed. Edinburgh: Edinburgh University Press

Meyer, Charles F. (2002): *English Corpus Linguistics. An Introduction*. Cambridge: Cambridge University Press

National Counterterrorism Center (ed.) (2009): *2008 Report on Terrorism*. Washington: Office of the Director of National Intelligence – https://www.fbi.gov/stats-services/publications/terror_08.pdf (11 April 2016)

Nuopponen, Anita (1996): "Concept Systems and Analysis of Special Language Texts. Towards a Terminological Analysis Method." Gerhard Budin (ed.): *Multilingualism in Specialist Communication. Proceedings of the 10th European LSP Symposium, Vienna.* Vol. II. Vienna: TermNet, 1069-1078

ÖNORM A 2704 (2015): *Terminologiearbeit – Grundsätze und Methoden*. Wien: Austrian Standards Institute

Teubert, Wolfgang (2007): "Language and Corpus Linguistics." Wolfgang Teubert, Anna Čermáková (eds): *Corpus Linguistics. A Short Introduction*. London: Continuum, 1-58

Teubert, Wolfgang; Anna Čermáková (2007): "Directions in Corpus Linguistics." Wolfgang Teubert, Anna Čermáková (eds): *Corpus Linguistics. A Short Introduction*. London: Continuum, 59-134

University of Georgia (n.d.): UGA Tobacco Documents Corpus – http://www.tobaccodocs.uga.edu/TDC/ (06 June 2016)

Author

Georg Löckinger holds a Master's degree and a PhD in translation studies from the University of Vienna. His award-winning doctoral thesis deals with translation-oriented reference tools and includes an empirical study on the information needs of professional translators. Since September 2013, Georg Löckinger has been a professor of technical communication at the University of Applied Sciences Upper Austria. Apart from his academic work, Georg Löckinger has extensive hands-on experience as a translator and terminologist in various professional settings. Furthermore, he is actively involved in national and international standardisation with regard to language technologies and language resource management.
E-mail: georg.loeckinger@fh-wels.at
Website: https://www.fh-ooe.at/en/wels-campus/studiengaenge/bachelor/product-design-and-technical-communication/

Neu bei Frank & Timme

TransÜD. Arbeiten zur Theorie und Praxis des Übersetzens und Dolmetschens

Herausgegeben von
Prof. Dr. Klaus-Dieter Baumann,
Dr. Susanne Hagemann,
Prof. Dr. Dr. h.c. Hartwig Kalverkämper,
Prof. Dr. Klaus Schubert

Susanne Hagemann: **Einführung in das translationswissenschaftliche Arbeiten.** Ein Lehr- und Übungsbuch. ISBN 978-3-7329-0125-8

Anja Maibaum: **Spielfilm-Synchronisation.** Eine translationskritische Analyse am Beispiel amerikanischer Historienfilme über den Zweiten Weltkrieg. ISBN 978-3-7329-0220-0

Franziska Heidrich: **Kommunikationsoptimierung im Fachübersetzungsprozess.** ISBN 978-3-7329-0262-0

FFF: Forum für Fachsprachen-Forschung

Herausgegeben von
Prof. Dr. Dr. h.c. Hartwig Kalverkämper

Eva Martha Eckkrammer: **Medizin für den Laien: Vom Pesttraktat zum digitalen Ratgebertext.** 2 Bände im Schuber. ISBN 978-3-86596-312-3

Kommunikation – Partizipation – Inklusion

Herausgegeben von
Dr. Bettina M. Bock,
Prof. Dr. Ulla Fix,
Daisy Lange

Nathalie Mälzer (Hg.): **Barrierefreie Kommunikation – Perspektiven aus Theorie und Praxis.** ISBN 978-3-7329-0231-6

TTT: Transkulturalität – Translation – Transfer

Herausgegeben von
Prof. Dr. Dörte Andres, Dr. Martina Behr,
Prof. Dr. Larisa Schippel,
Dr. Cornelia Zwischenberger

Marc Orlando: **Training 21st century translators and interpreters: At the crossroads of practice, research and pedagogy.** ISBN 978-3-7329-0245-3

Christian Trollmann: **Nationalsozialismus auf Japanisch?** Deutsch-Japanische Beziehungen 1933–1945 aus translationssoziologischer Sicht. ISBN 978-3-7329-0281-1

Frank & Timme
Verlag für wissenschaftliche Literatur

Wittelsbacherstraße 27a, D-10707 Berlin
Telefon (030) 88 66 79 11, Fax (030) 88 39 87 31
info@frank-timme.de, www.frank-timme.de