Anne Kimmes (Heidelberg) and Hilko Koopman (Mannheim)

COLLOCATION | ANALYZER –

An Electronic Tool for Collocation Retrieval and Verification

Publisher:

Wissenschaftlicher Verlag Trier

 ISSN: 2191-1916

Anne Kimmes (Heidelberg) and Hilko Koopman (Mannheim)

COLLOCATION │ ANALYZER –

An Electronic Tool for Collocation Retrieval and Verification

Abstract:

Collocations cause considerable difficulties to anyone who wants to produce texts in a foreign language. Based on theories of collocations developed by John Rupert Firth and Franz Josef Hausmann, we will show that the current reference works do not satisfactorily provide help in the field of collocations. For this reason, we developed an application called COLLOCATION │ANALYZER to fill this market gap with a user-friendly electronic tool for collocation retrieval and verification. COLLOCATION│ ANALYZER allows the user to perform various tasks: retrieving unknown collocations; verifying tentatively available collocations; comparing collocations of semantically similar bases; viewing contextual examples; sorting and filtering the results in a number of different ways; as well as looking up words in the renowned *Merriam-Webster* online dictionary and thesaurus directly from within the application.

Kollokationen bereiten all jenen erhebliche Schwierigkeiten, die Texte in einer Fremdsprache verfassen möchten. Gestützt auf den Kollokationstheorien von John Rupert Firth und Franz Josef Hausmann wird gezeigt, dass die derzeit verfügbaren Hilfsmittel im Bereich des gebundenen Sprachgebrauchs nur unzureichend Hilfe bieten. Mit der Absicht, diese Marktlücke zu schließen, entstand der COLLOCATION │ ANALYZER, ein neues benutzerfreundliches elektronisches Hilfsmittel zum Nachschlagen und zur Überprüfung von Kollokation. Mit Hilfe des COLLOCATION│ ANALYZER können unbekannte Kollokationen gesucht werden; tentativ verfügbare Kollokationen können überprüft werden; der Nutzer kann Verwendungsbeispiele in ihrem Kontext abrufen; er kann die Ergebnisse auf unterschiedliche Weise sortieren und filtern; sowie unbekannte Wörter direkt im renommierten *Merriam-Webster* Online-Wörterbuch und Thesaurus nachschlagen.

Keywords:

collocations; word combinations; British Contextualism, London School of Linguistics, translation; electronic dictionary; electronic reference tool; corpus.

Kollokationen; Wortkombinationen; Britischer Kontextualismus; London School of Linguistics; Übersetzung; elektronisches Wörterbuch; Hilfsmittel; Korpus.

Contents:

1 Introduction 2
2 Theories of Collocations 3
3 Collocations and Text Production in Foreign Languages 5
4 Reference Works in the Field of Collocations 8
5 COLLOCATION | ANALYZER 10
5.1 Make-up and Contents 10
5.2 Start 12
5.3 Collocation Finder 14
5.4 Collocation Comparison 21
5.5 Cluster Comparison 23
5.6 Sources 25
6 Summary and Concluding Remarks 27
7 References 28

1 Introduction

Collocations are conventional and recurrent lexical chunks consisting of two words.
The *Longman Dictionary of Contemporary English* simply defines the term *collocation* as "the way in which some words are often used together, or a particular combination of words used in this way" and provides the example "to commit a crime" to illustrate this concept. A criminal can *commit*, *carry out* or perhaps *perpetrate a crime*. But it is very unidiomatic to say *to make a crime* or *to do a crime*.

Collocations are ubiquitous in our everyday language. We use them naturally without paying particular attention to them. In the following passage, collocations are printed in italics to demonstrate how widespread they are. The examples are taken from the *Oxford Collocations Dictionary for Students of English* (Study Page 12) and have been slightly modified.

After *committing the crime*, the *defendant was arrested* and *remanded in custody to await trial*. He was *charged with threatening behavior*, an offense that *carries a sentence* of up to

two years in jail. The jury *returned a verdict of guilty* and the judge *passed sentence*. The man was *found guilty* and was s*ent to prison*.

Collocations are a serious stumbling block for anyone who wants to produce texts in a foreign language. For this reason, foreign language teaching has increasingly focused on collocations, or phraseology in general, in the past years. However, collocations not only create difficulties for students of a foreign language but also for language professionals such as translators or interpreters. Despite the awareness of collocations as a typical source of error, there are no reference works available that satisfy the foreign language user's needs.

In this article, we will briefly present different theories of collocations (Section2), explain why collocations are a problematic area in foreign language text production (Section 3), and present the most important reference works in the field of collocations (Section 4). Section 5 is the heart of this article. We will introduce the reader to COLLOCATION | ANALYZER, a new electronic tool for collocation retrieval and verification, that is meant to fill the market gap with a user-friendly reference work in the field of collocations.

2 Theories of Collocations

Different scholars at different times have concentrated on various aspects of word combinations and have come forth with a variety of contrasting definitions. The most prominent definition may have been developed by John Rupert Firth, the founder of the London School of Linguistics and a representative of British Contextualism. Firth was one of the first linguists to deal with collocations and put forth a theory of *meaning by collocation* which suggested that the meaning of a word can only be conveyed when the other words with which it usually co-occurs are also considered: "One of the meanings of night is its collocability with *dark*, and of *dark*, of course, collocation with *night*" (Firth 1957: 196). In his understanding of collocation, Firth emphasized the importance of "know[ing] a word by the company it keeps" (1957: 179). Firth largely influenced British linguistics. However, the British Contextualism represented by Firth and his students (sometimes also called neo-Firthians) mainly focused on collocations as the statistically measurable likelihood of two items to co-occur.

In the 1980s, the German linguist Franz Josef Hausmann offered a new view of collocations that proved to be very applicable to foreign text production or translation in particular. In his theory of word combinations (see Figure 1), Hausmann differentiates between fixed and non-fixed word combinations.

Collocations fall into the category of affine, non-fixed combinations and are used very commonly in every language. Counter creations such as *eyes wide shut* or *black milk* are rarely used and are striking to the reader or hearer. Free combinations—so called co-creations in Hausmann's terminology—make up the largest part of word combinations and are not particularly noticed.

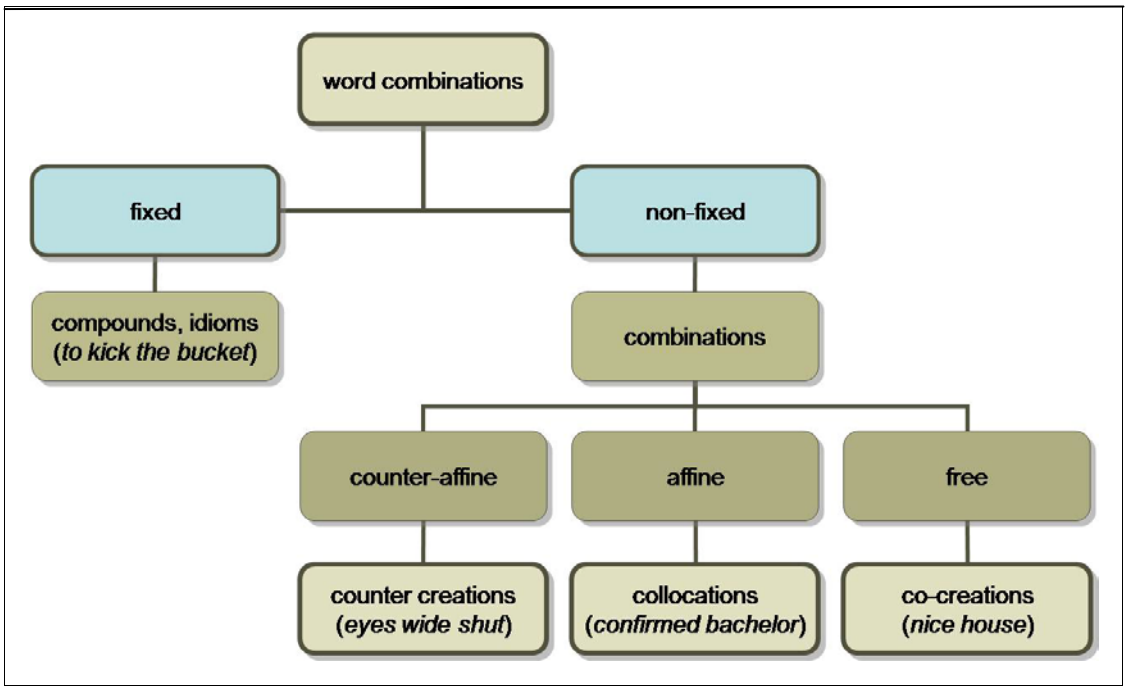

Fig. 1: Hausmann's Theory of Word Combinations (Hausmann 1984: 399)

Hausmann also introduced a new understanding of the two constituents of a collocation. British Contextualism distinguished between the *node* and the *collocate* of a collocation with the node being the part of the combination that was currently being studied. In this way, each word of the collocation could be either node or collocate depending on the research interest. In contrast to that, Hausmann speaks of a *base* and a *collocator*. These two

constituents of a collocation bear a hierarchical relationship: The base determines the collocator and the collocation as a whole. The collocator is determined by the base. Hausmann presented six different types of collocations (see Table 1). In four of the six types, the base (printed in bold) is represented by a noun.

1. verb + **noun** (object)	t*o fall in love*, *to express admiration*
2. adjective + **noun**	*broad implications*, *serious consequences*
3. **noun** (subject) + verb	*a problem persists*, *a complication arises*
4. noun + (preposition) + **noun**	*career goal*, *job market*
5. adverb + **adjective**	*deadly serious*, *achingly funny*
6. **verb** + adverb	*to sleep soundly*, *to work hard*

Table 1: Hausmann's Six Different Types of Collocations (Hausmann 1999: vii)

3 Collocations and Text Production in Foreign Languages

Hausmann's theory is particularly applicable to (foreign) text production because it depicts the process of finding an appropriate equivalent in the target language. Native speakers of a language generally use collocations intuitively. The speaker or writer simply puts the base and an appropriate collocator together in the sentence without having to reflect on his or her particular word choices. However, collocations do present a serious obstacle whenever a text is to be produced in a foreign language. Collocations differ from one language to another and in doing so also reflect the culture of a country or language area. In English, for instance, *to make coffee* is the correct collocation. The French *faire du café* and the Spanish *hacer café* work with the same verb. 'Faire' (FR) and 'hacer' (SP) also mean 'to make.' In German and Dutch, the corresponding collocations *Kaffee machen* and *coffie maken* ('machen' / 'maken' = 'to make') are also correct. In German, however, *Kaffee kochen* ('kochen' = 'to cook') is a far more common expression and in Dutch (DU), *koffie zetten* ('zetten' = 'to set / to put') is preferably used.

If a native speaker of German, for instance, was to translate the collocation *Kaffee kochen* into the four other languages he or she might be tempted to translate the collocation literally and say **to cook coffee* (EN), **(faire) bouillir du café* (FR), **(hacer) hervir café* (SP), and **koffie koken* (DU).

The following table summarizes the literal translations of the German collocation *Kaffee kochen* and the correct equivalents in English, French, Spanish, and Dutch. In all of the cases, the literal translation is wrong.

	Literal translation / correct collocation			
German	**English**	**French**	**Spanish**	**Dutch**
Kaffee kochen	**to cook coffee* *to make coffee*	**(faire) bouillir du café* *faire du café*	**(hacer) hervir café* *hacer café*	**koffie koken* *koffie zetten*

Table 2: The German collocation *Kaffee kochen* and its equivalents in English, French, Spanish, and Dutch

A key characteristic of collocations is that their constituents can only be combined with a limited number of other words. The translator—or anyone who needs to produce a text in a foreign language—must be familiar with these combinatory rules. Native speakers of the languages referred to above would probably understand all of the combinations represented with an asterisk in Table 2, but they would consider them unidiomatic. Whenever a language professional such as a translator or interpreter (unintentionally) does not comply with these combinatory rules, the quality of his complete work is questioned. It is unnecessary to mention that this may have disastrous consequences for someone who earns his keep by producing text in a foreign language.

For this reason, everyone, and language professionals in particular, must be aware of the fact that the collocation as a whole must be transferred into the foreign language. The base of a collocation, which is generally a noun (see Table 1), is usually known or can be easily retrieved with the help of bilingual dictionaries. If a German translator wishes to translate *Kaffee kochen* into English, he or she is likely to know the English equivalent of 'Kaffee,' namely 'coffee.' It is the translation of the collocator 'kochen' that creates difficulties. As has

been seen before, the literal translation **to cook coffee* is not a good choice in this situation. He or she needs to find an adequate English verb that can be idiomatically combined with the noun 'coffee' and that carries the meaning of the German phrase *Kaffee kochen*. Thus, the difficulty in collocations lies not in translating the base but in finding an appropriate collocator in the target language.

The translation of a collocation is therefore a process involving two distinct steps. First, the translator must choose an item in the paradigmatic dimension that reflects all of the characteristics of the base in the source language (see Diller and Kornelius 1978: 34-35). Even with a noun as simple as 'coffee' this might not be an easy task, especially given the fact that collocations also reflect different cultures and customs. The translator must be aware that the conception of 'coffee' fundamentally differs from one country to another and that the word choice must be adapted accordingly. The Italian 'caffè,' the Spanish and the French 'café' do translate to the English word 'coffee' but the concept behind the word is not necessarily the same. If an Italian orders a 'caffè' in a restaurant, he or she is likely to receive what an anglophone person would call an 'espresso.' While for an Italian, a regular coffee is a small cup of espresso, for the English-speaking world, a regular coffee is a larger cup of filter coffee. This shows that a translator is not only a mediator between different languages but also between different cultures. A professional translator therefore requires a sound knowledge and detailed familiarity of both the language and the culture of the target area for which he or she translates.

Once an item in the target language possessing all semantic characteristics of the base in the source language has been located, the translator needs to make a selection in the syntagmatic dimension. This is the second selection process and involves finding an appropriate collocator in the target language that can be idiomatically combined with the base (see Diller and Kornelius 1978: 36-39). The translation of the collocator is independent of the collocator in the source language. While, in general, equivalences such as German 'Kaffee' = English 'coffee' can be established for base words, this is not possible for collocators. When looking at the collocators separately, the equivalence of the German word 'kochen' would be 'to cook' in English. But 'kochen' as in the collocation *Kaffee kochen* must

be translated as 'to make' or maybe as 'to prepare' (*to make coffee* / *to prepare coffee*). Every linguistic element of every language can be combined with a limited number of other elements, whereas this combinability is arbitrary, not complying with any linguistic rules and difficult to depict. The translation of collocations therefore involves constant compatibility checks in the syntagmatic dimension.

4 Reference Works in the Field of Collocations

It is impossible for a translator to know all collocations of a language. The number of collocations is simply too large to be learnt and internalized entirely (Angelone and Connelly 2007: 28, 31). But despite this virtually infinite number of collocations, translators are often highly qualified professionals who, in general, already have a possible collocator in the target language in mind. Translators therefore need first-class reference works that allow them to retrieve a collocation they do not know and to verify if their tentative solution is a correct collocation.

However, up to now, there is no reference work available that satisfactorily assists the translator in finding a collocation in the foreign language (see Bahns 1996; Butina-Koller 2005; Kornelius 1995a; Steinbügel 2005). Dictionaries (be they monolingual or bilingual dictionaries, learner's dictionaries for encoding purposes or general dictionaries for decoding purposes) often list the collocations in the wrong entries. Collocations are only of use in the dictionaries when they are listed at the entry of the base. A listing of the collocation at the entry for the collocator only helps if the translator also has a collocation in mind and wants to verify it. No anglophone person would look under the entry of the German verb 'kochen' ('to cook') if he or she were to translate *to make coffee* into German. The collocation would have to be listed under the base entry 'coffee' / 'Kaffee.'

The number of specialized collocation dictionaries is very limited. For the English language, there are only two books that are worth mentioning:

- The *Oxford Collocations Dictionary for Students of English* (2002), which—according to the book's cover—contains 150,000 collocations of 9,000 nouns, verbs and adjectives as well as over 50,000 examples of collocations in context.
- The *Student's Dictionary of Collocations* (1999), which is based on the *BBI Dictionary of English Word Combinations* (rev. ed. 1997) and contains 18,000 main entries and 90,000 collocations according to the editors.

Both of these books are useful for any non-native speaker of English who wishes to produce text in English. However, their main disadvantage is that they cannot portray the language in its entirety. Print dictionaries are always restricted in their size. Another disadvantage is, of course, that print dictionaries are not very user-friendly. Modern translators carry out almost all of their daily work on the computer: Translation jobs come in and are delivered via e-mail. Texts are processed in word processing software such as *Microsoft Word* and are translated, most of the time, with the help of computer-aided translation tools such as *SDL Trados*. Communication takes place via instant messaging or software such as *Skype* that allows users to make telephone calls over the Internet and use features such as file transfer and video conferencing. Given this work environment, print dictionaries are somewhat outdated. Modern translators have all of their reference works readily available on their computers, either as software installed from a CD or directly online. In this way, unknown words and phrases can be looked up within seconds, cross-references to other entries can be followed directly and no time is wasted flipping through the paper pages of a book. After all, translation is a service in which the motto "time is money" fully applies. Translations are generally charged by the number of words or lines in a text, which implies that the more time a translator needs to do the necessary research the less money he or she earns in an hour. It would therefore be desirable for the collocation dictionaries also to become available in electronic form.

Another major drawback of the two collocation dictionaries mentioned above is that they do not fully disclose their sources. The "main source" used to retrieve the collocations listed in the *Oxford Collocations Dictionary for Students of English* is the 100-million-words *British*

National Corpus. According to the publishers, other sources are also exploited but they are not specified (Crowther, Dignen and Lea 2002: viii). The *Student's Dictionary of Collocations* does not reveal its sources at all. It is in no way clear where the collocations listed in this book come from, which makes the information of course not very reliable.

An extensive and fully reliable electronic reference tool in the field of collocations is urgently needed on the market. Various linguists are aware of this shortcoming and have developed different outlines and concepts of such an electronic reference work in the field of collocations. As early as 1995, Kornelius offered suggestions for a corpus-based electronic collocation dictionary with various retrieval options ("Vom Printwörterbuch zum elektronischen Kollokationswörterbuch"). In 2007, Jehle outlined a collocation dictionary on DVD and Angelone presented an *E-Collocation Trainer* that can be used to store, retrieve and verify collocations. All of these models have different advantages and drawbacks, but they all have one problem in common: they have not been implemented yet. In 2003, Holderbaum presented a bilingual collocation database (English/German) that is designed in a way that satisfies most of the needs of foreign language users. The database is available online and therefore does not face problems of limited size or long access times. Unfortunately, however, this database hardly contains any entries and is therefore of little practical use for the time being.

In the following, COLLOCATION │ANALYZER will be presented. It is a fully functional electronic tool for collocation retrieval and verification that is meant to fill the current gap and facilitate the translator's daily work.

5 COLLOCATION │ANALYZER

5.1 Make-up and Contents

COLLOCATION │ANALYZER was designed out of the need for an electronic tool that translators can use to quickly retrieve, verify, and compare collocations. It is based on *Microsoft Excel* and is optimized for *Excel 2007*. COLLOCATION │ANALYZER is implemented with *Microsoft Visual Basic for Applications* (VBA), which is built into most *Microsoft Office*

applications to enable developers to create custom solutions with the programming language *Microsoft Visual Basic*. In this way, VBA can automate and extend the functionality of *Microsoft Office* applications.

Up to now, COLLOCATION│ANALYZER contains collocations of 48 noun bases, belonging to four distinct semantic fields:

Love
admiration, adoration, affection, ardor, devotion, fervor, fondness, infatuation, love, passion, rapture, tenderness
Result
conclusion, consequence, corollary, effect, implication, outcome, outgrowth, ramification, repercussion, result, score, upshot
Trouble
complication, difficulty, dilemma, distress, hardship, hindrance, obstacle, plight, predicament, problem, quandary, trouble
Work
assignment, career, employment, job, livelihood, occupation, position, post, profession, task, vocation, work

Table 3: Semantic Fields and Noun Bases in the COLLOCATION │ANALYZER

The collocations that are available in the tool have been retrieved from three different sources. The first two are both of the print collocation dictionaries *Oxford Collocations Dictionary for Students of English* and the *Student's Dictionary of Collocations*. The largest group of collocations, however, was manually retrieved from a text corpus consisting of all the news writing published in the online edition of the renowned *Washington Post* newspaper between March 1, 2006 and February 28, 2007. The corpus contains a total of more than 37 million words that appeared in 68,122 different articles within nine different newspaper sections (*A Section*, *Editorial*, *Financial*, *Health Tab*, *Metro*, *Sports*, *Style*,

Weekend, *Weekly Virginia*). The corpus was searched for collocations of each of the 48 base words until either the entire corpus (37,642,155 words) was scanned or at least one hundred different word combinations were retrieved. The collocations were then transferred with all of their context information into a complex *Excel* database.

The collocation itself (e.g. *to look for a job*), and the entire sentence in which the collocation occurred was copied into the *Excel* database. Each time, it was noted to which semantic field and base word the collocation belongs. Moreover, the source of the collocation was recorded (*Washington Post*, *Oxford Collocations Dictionary for Students of English* or *Student's Dictionary of Collocations*). For collocations retrieved from the *Washington Post*, the newspaper section and the month and year in which the collocation appeared in the article was noted. For each collocation, the collocation type (see Table 1) was determined and marked in the *Excel* database. The database was equipped with a number of filters, drop-down menus and standard lists that facilitated the handling of the large amount of collocational data. Macros were also employed to automate parts of the data entering process. For instance, for each collocation a "reduced collocation" was automatically entered into the database, in which the base was replaced by a tilde (~). In the reduced collocation form, (grammatical) characteristics such as the use of an article or a plural form were disregarded. This was necessary to make collocations such as *to look for **a** job* and *to look for work* fully comparable.

COLLOCATION │ANALYZER is a VBA implementation of the described *Excel* database by means of which all of the contents of the database can be easily and comfortably queried. The tool consists of five different tabs in each of which different functions are available to retrieve, verify and compare collocations. In the following sections, each of the tabs will be described in further detail.

5.2 Start

Upon starting COLLOCATION│ANALYZER, the user is presented with the following welcome or start screen (Figure 2), which serves as an overview of the collocation tool. At the top of

the screen, next to the application's logo, COLLOCATION | ANALYZER is described in a few words.

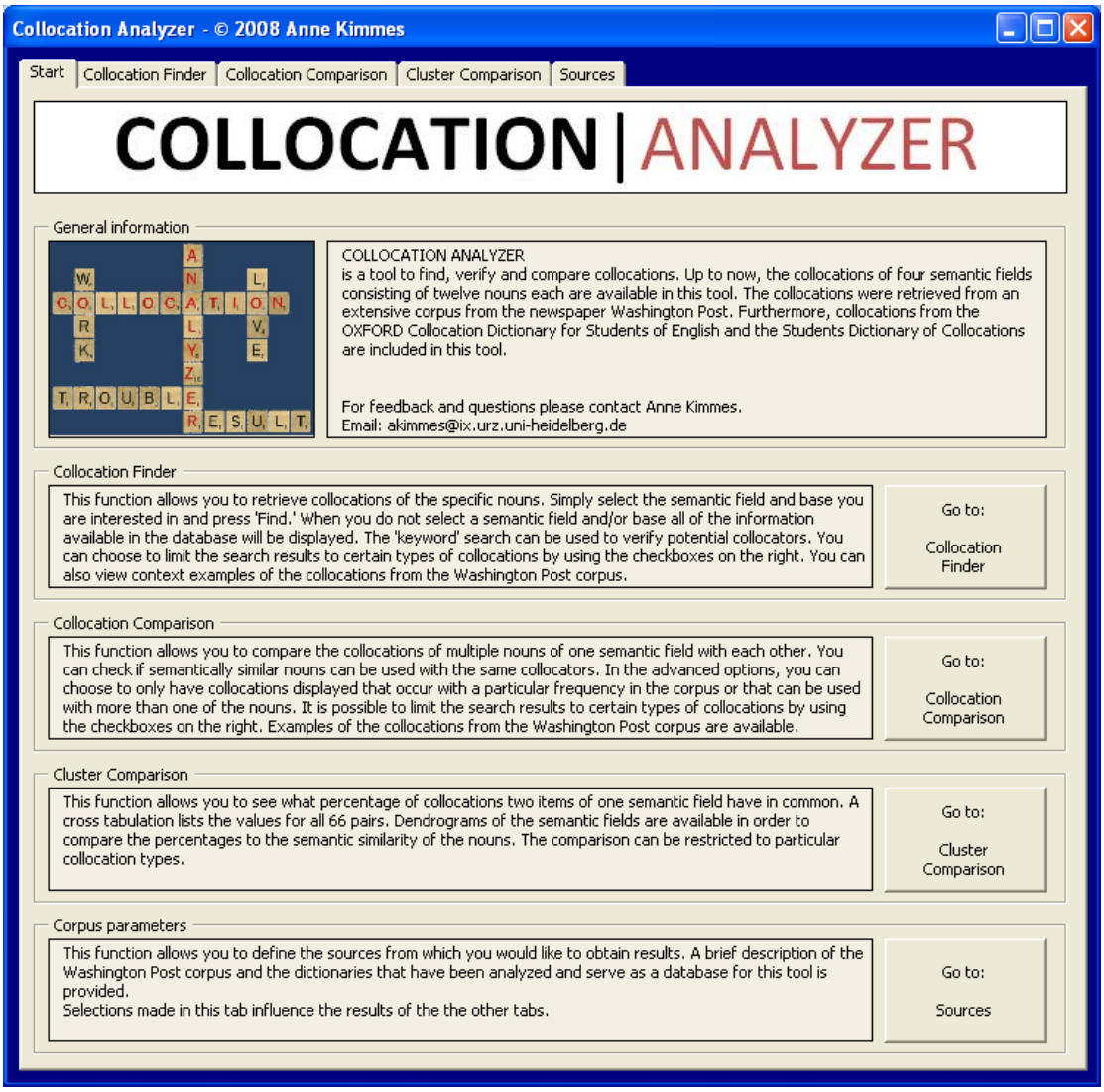

Fig. 2: Tab 1 of COLLOCATION | ANALYZER: Start

Moreover, there is a little box for each of the other four tabs in which the functions available in each particular tab are briefly explained. The user can access the tabs either by clicking on the button next to the short explanation or simply by clicking directly on the desired tab at the top of the screen.

5.3 Collocation Finder

The Collocation Finder tab (Figure 3) is the heart of the COLLOCATION | ANALYZER. It allows users to retrieve and verify collocations of the 48 noun bases currently available in the tool.

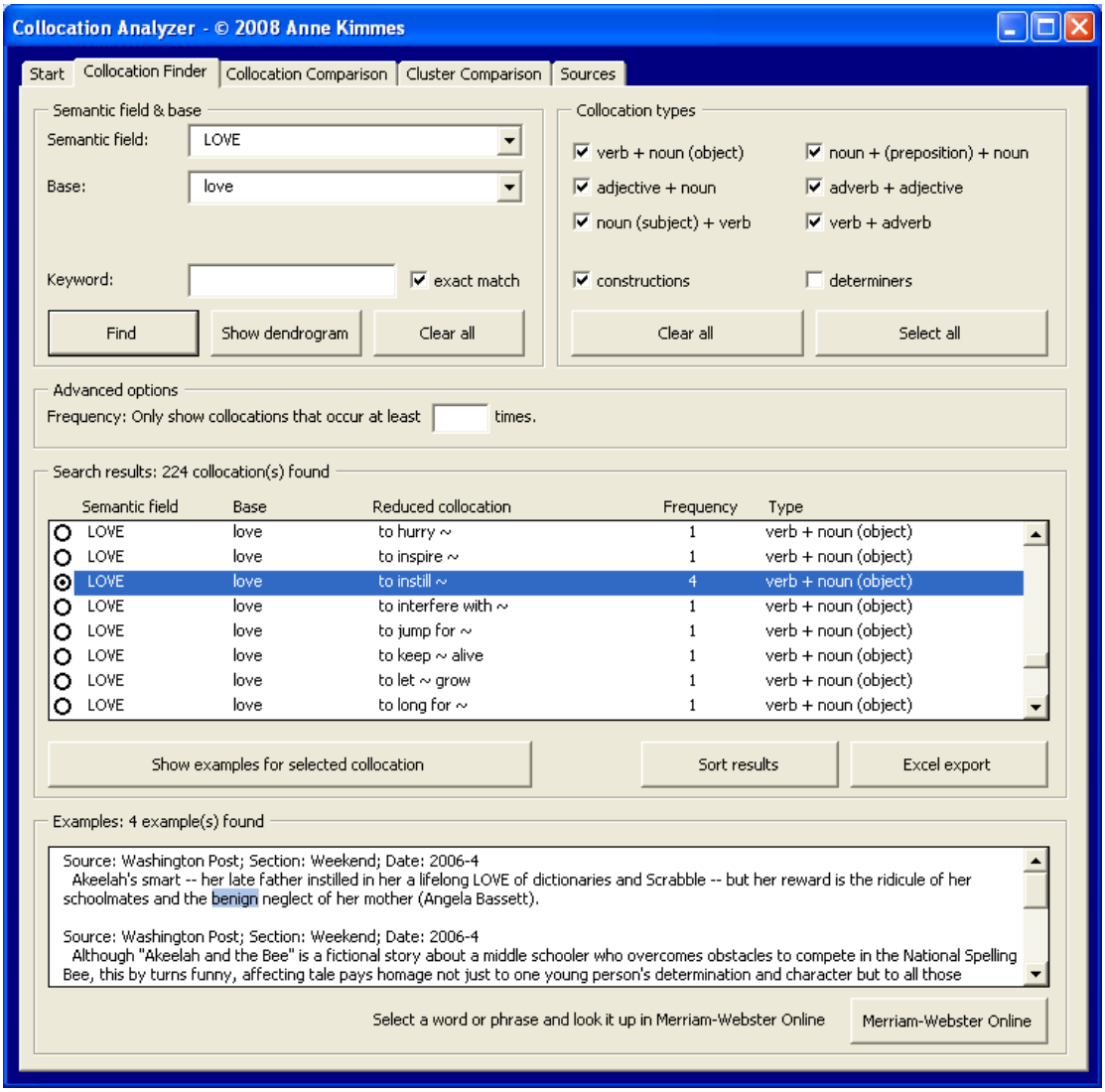

Fig. 3: Tab 2 of COLLOCATION | ANALYZER: Collocation Finder

In order to RETRIEVE unknown collocations, the user must select the semantic field and base he or she is interested in from the pull-down menus at the upper left-hand side of the interface and press Find.

The Keyword search can be used in order to VERIFY tentatively available collocations. The user can also select the semantic field and base he or she is interested in from the pull-down menus, enter the potential collocator in the keyword box and press Find. If the checkbox Exact match is marked, only collocations containing the exact keyword are displayed. If the checkbox is deactivated, results will also be displayed that contain the keyword somewhere in the collocator. For instance, a keyword search for "rea" within the semantic field LOVE and base "love" without checking the Exact match box will result in collocations such as *love breaks something*, *great love*, *real love*).

When a semantic field is selected from the first pull-down menu, the twelve nouns belonging to that field are available in the second menu. When no semantic field and no base are selected, all of the information available in the database will be displayed. When the user selects only a semantic field but no base within the field, the collocations of all twelve nouns of that field will be displayed. For instance, if the user selects the semantic field LOVE and then clicks on Find, all of the collocations of all twelve bases within the semantic field are displayed.

The same is true for the collocation verification option via the keyword search. In general, the user should select the semantic field and base and enter a collocator in the keyword field in order to verify if the collocator is available for the selected base. However, it is also possible to leave the semantic field and base menus blank and perform the keyword search without a prior selection. In this way, all the collocations that contain the keyword are displayed. For instance, a keyword search for "real" within the semantic field LOVE with the checkbox Exact match activated leads to the collocations *real affection*, *real love*, and *real passion*. A search for the keyword "real" in all four semantic fields (no semantic field selected) reveals the additional collocations *real effect*, *real repercussion* (RESULT), *real difficulty*, *real dilemma*, *real distress*, *real hardship*, *real problem*, *real trouble* (TROUBLE), as well as *real career*, and *real vocation* (WORK).

At the upper right-hand side of the interface, the user can further specify the collocation types he or she is interested in. The six collocation types established by Hausmann (see Table 1) are available as well as constructions and determiners. These categories were

added because foreign language users also need grammatical collocations (so-called "colligations") to assure idiomatic expression in the target language. Even though constructions and determiners do not fall within the narrower definition of collocations, they nevertheless provide the user with valuable information regarding the usage of the nouns. In the same way, COLLOCATION│ANALYZER also contains word combinations that can be regarded as free combinations. It is difficult, if not impossible, to draw a clear line between collocations and free combinations (so-called co-creations, see Figure 1). In any case, these definitions are of little value for the foreign language user as he or she is only interested in correct and idiomatic expression. For this reason, all of the word combinations are included in the tool. By default, the checkboxes of all collocation types except Determiners are selected at the upper right-hand side of the interface.

An error message pops up if none of the checkboxes for the collocation types is marked. COLLOCATION│ANALYZER contains a number of error messages that appear whenever the user does not make the necessary selections. The tool can be used intuitively, but whenever the user is on the wrong path, the error messages tell him or her what to do in order to perform the desired retrieval, verification or comparison task.

By clicking on the buttons Clear all and Select all, the user can check or uncheck all collocation types at once. The Semantic field and base selection on the left-hand side also contains a button to Clear all selections made in this part of the interface.

The collocations are displayed in the Search results pane in the middle of the interface. In Figure 3, all of the collocations of the noun base "love" within the semantic field LOVE that belong to either one of the six collocation types or that can be defined as a construction are displayed. The user can see at the top of the search results pane how many collocations with the corresponding criteria were found ("224" in the example in Figure 3). The search results pane indicates the semantic field and the base each collocation belongs to, the reduced collocation, the frequency (meaning the number of times this particular collocation was retrieved from the selected sources) and the collocation type.

One must bear in mind that an identical collocation can be available several times within the tool. It should be noted, however, that each time a collocation was retrieved from the *Washington Post* corpus it was also entered into the *Excel* database underlying COLLOCATION│ANALYZER. Collocations from one of the dictionaries were only entered once into the database. The frequency of collocations from the newspaper corpus therefore indicates how often a collocation is used. When a collocation was listed in one of the dictionaries, it only indicates that this collocation can be used, but it provides no information on how common the collocation is.

The Advanced options above the search results pane allow the user to only have collocations displayed that occurred more often than a particular number of times. In order to assure that very uncommon collocations are not shown, the user can enter any number higher than "1" in the box. Even though frequency alone does not give information about the affinity of a collocation, it may guide the user in his word choice.

The search results can also be sorted in various ways to change the alphabetical sorting by default. It is particularly useful to sort the results according to their frequency in the order Z to A (highest frequency first). In this way, the user can scroll through the results and see which collocations were very commonly used in the corpus and which were not.

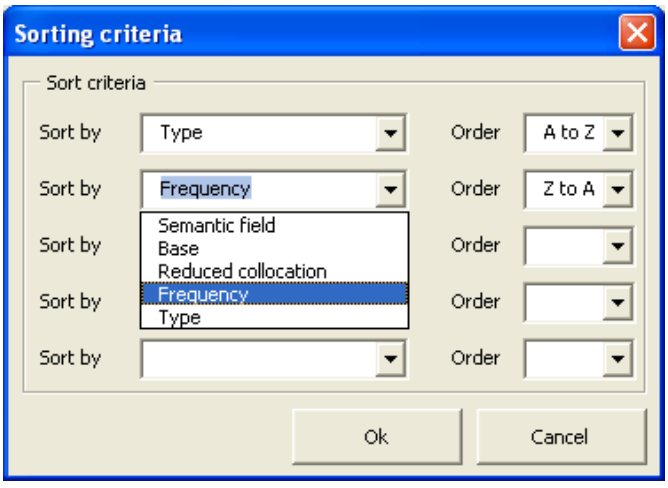

Fig. 4: Sorting Options Within Tab 2 of COLLOCATION│ANALYZER

If the user would like to work with the data in an external application, he or she can export the search results into *Microsoft Excel* with the corresponding button underneath the search results pane.

As has been said before, COLLOCATION│ANALYZER also contains full contextual examples of all collocations. The user can either double click on a collocation in the search results pane, or he or she can select a collocation with the mouse and click on the button Show examples for selected collocation. As a result, the contextual examples are displayed in the examples pane at the bottom of the interface. The user can see at the top of the pane how many examples were found for the selected collocation ("4" in the example in Figure 3). The source of the collocation is indicated above each example. In the case of the *Washington Post*, the newspaper section and the date of publication are also provided. The example generally comprises the entire sentence in which the collocation was naturally used in the newspaper article. The base word is displayed in CAPITAL LETTERS in the example sentence to catch the user's eye right away. The collocation dictionaries do not always contain a contextual example for the collocation. In this case, the message "Sorry, this dictionary does not contain an example for this collocation." will be displayed after the source (e.g. Source: OXFORD Collocations). Sometimes the dictionaries do not give an example for the collocation but differentiate between the different meanings of polysemous words. Of course, only the collocations belonging to the meaning of the given semantic field were entered in the *Excel* database underlying COLLOCATION│ ANALYZER.

While the collocations from the search results pane can be exported to *Microsoft Excel* via a special button, the contents of the examples pane can be easily transferred to other applications via copy and paste.

COLLOCATION│ANALYZER also contains a direct link to the well-known *Merriam-Webster* online dictionary and thesaurus. Thus, it is possible to conveniently look up any unknown word directly from within the application. The user can select and highlight any word or phrase in the examples pane and click on the button Merriam-Webster Online (see Figure 3). A pop-up window appears in which the highlighted word from the examples pane is already

entered. The user can choose if he or she would like to look up the word in the dictionary or the thesaurus and click Search.

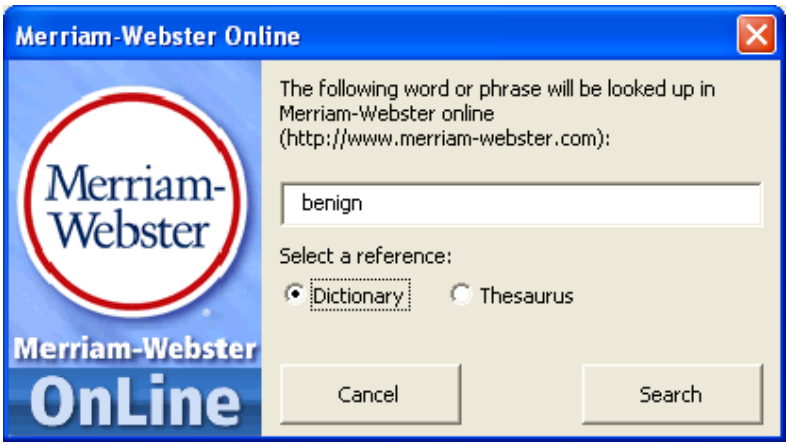

Fig. 5: Pop-up Window Directing User to Merriam-Webster Online

As a result, the search query is performed and automatically opened in the user's default web browser. The user can click on the button Merriam-Webster Online at any time while using COLLOCATION | ANALYZER and enter a word manually in the pop-up.

A final feature in the Semantic field and base section of the interface is the command Show dendrogram. COLLOCATION | ANALYZER contains dendrograms (also called tree diagrams) for the four semantic fields currently available in the tool.

A total of 330 native speakers of American English were asked to sort the members of the four semantic fields into groups according to their semantic similarity. Afterwards, a hierarchical cluster analysis was performed on the results (see Kimmes 2009). A dendrogram is a graphical representation of the hierarchical cluster analysis in which each noun is represented on a branch. The more native speakers sorted two items into the same group, the closer these two items are in the dendrogram. This means that the shorter the distance between two items is, the more semantic features those two items have in common.

When a semantic field is selected in the pull-down menu, only the dendrogram for this particular semantic field is displayed. If no field is selected, all four dendrograms can be accessed via the tabs at the top of the pop-up. In addition to the dendrograms, there is a tab

entitled Information in which explanations are given on how the dendrograms should be interpreted.

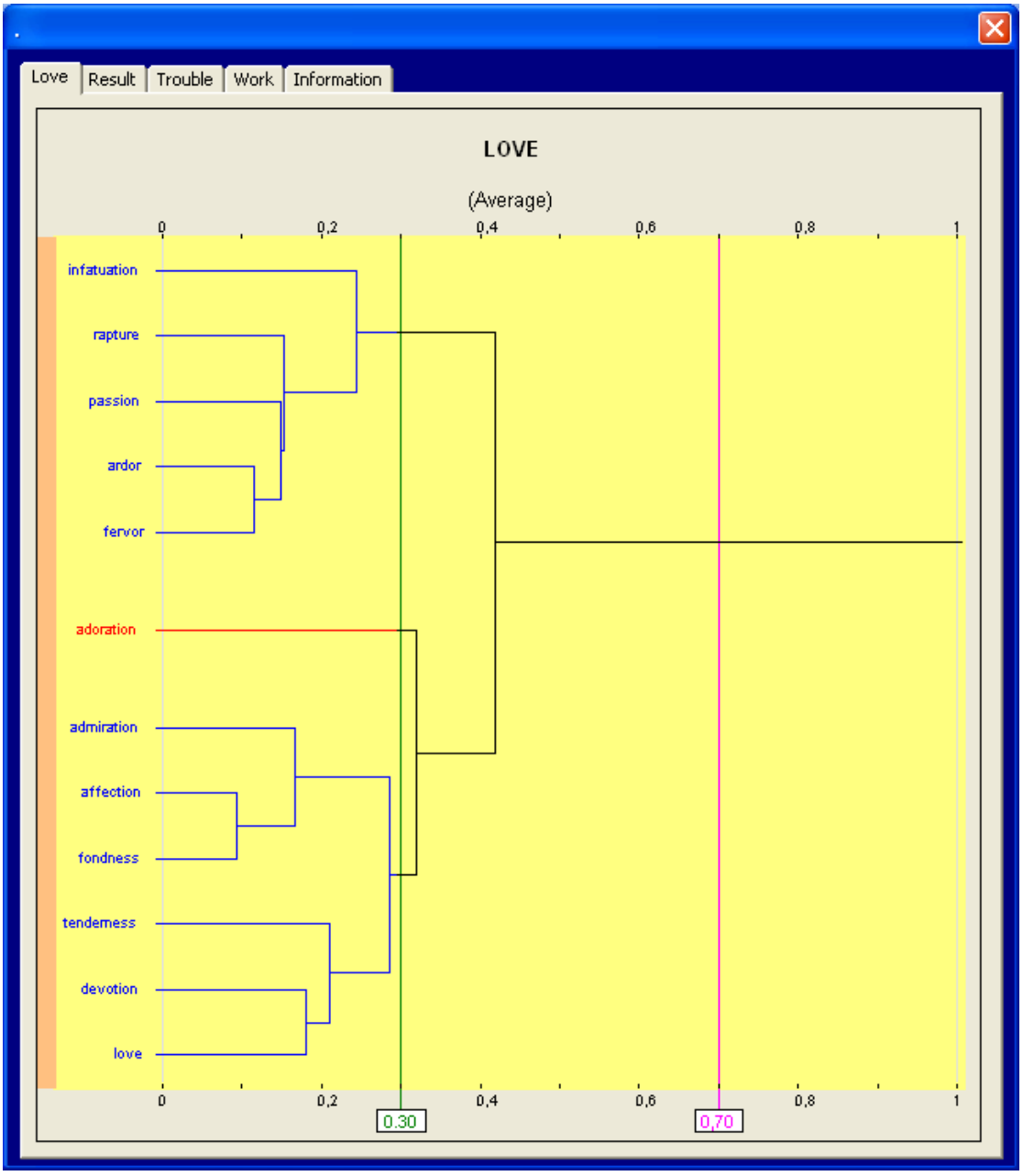

Fig. 6: Dendrograms Within the COLLOCATION | ANALYZER

As shown in Figure 6, "affection" and "fondness" are the two items of the semantic field LOVE that are closest in meaning according to native speakers questioned.

The dendrograms are of particular interest in tabs 3 and 4 of the COLLOCATION | ANALYZER when the collocations and clusters are compared. The collocational data in the following tabs in combination with the dendrograms can provide insight into possible relationships between the semantic similarity of two nouns and their collocational ranges.

5.4 Collocation Comparison

In the tab entitled Collocation Comparison, the user can compare the collocations of multiple nouns of one semantic field with each other. For this purpose, it is necessary to select a semantic field and at least two different bases from the pull-down menus at the upper left-hand side of the interface and click Compare. Optionally, the user can, once again, define the collocation types via the checkboxes at the upper right-hand side.

In the results pane, all of the collocations that fulfill the selected criteria are displayed. In the example in Figure 7, all of the adjective + noun collocations of the bases "consequence," "result," and "outcome" within the semantic field RESULT are shown.

There is a column for each of the selected bases in the results pane. The number indicates how many times the selected collocator was retrieved from the sources for the given base. In the example in Figure 7, the corpus and the dictionaries provided no examples for *favorable consequence*, one instance of *favorable result* and two occurrences of *favorable outcome*. The total Frequency is therefore three (0 + 1 + 2). The Match indicates for how many of the three bases the selected collocator was available in the sources. In the case of "favorable" in Figure 7, the Match is two. The sources provided examples for "favorable" with two of the selected bases: *favorable result* and *favorable outcome*.

In the Advanced options, the user can filter according to Frequency and Match. He or she can choose to only have collocations displayed that occurred at least a particular number of times (Frequency). In this way, uncommon combinations will not be shown. Moreover, he or she can choose to only see collocators that were found with more than one of the bases (Match).

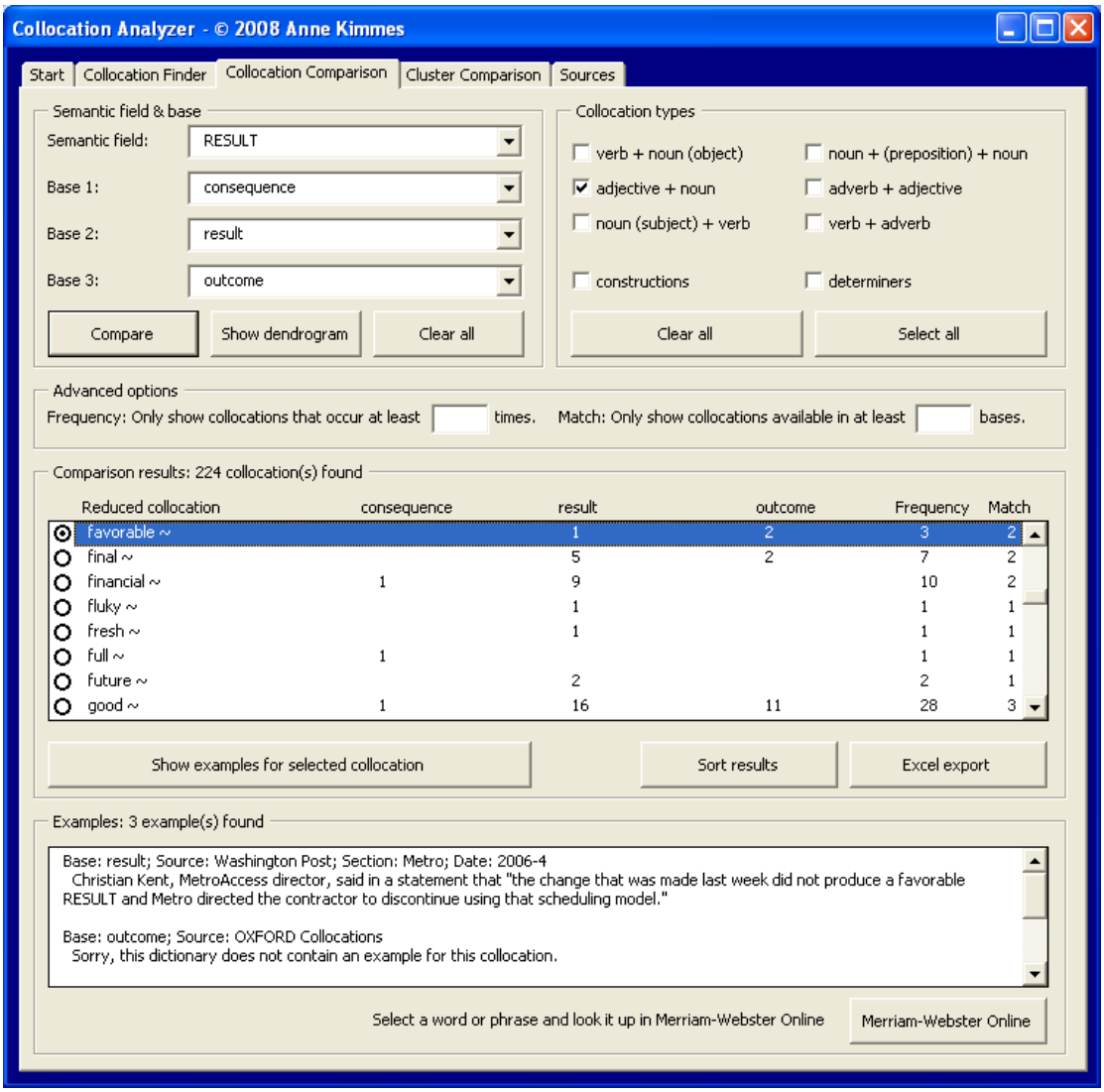

Fig. 7: Tab 3 of COLLOCATION | ANALYZER: Collocation Comparison

All of the other options are basically equal to the ones presented for the second tab (Collocation Finder): It is possible to sort the examples according to all of the columns available in the comparison results pane: the reduced collocation, match, frequency, as well as the three selected bases. A pop-up similar to the one presented in Figure 4 is available. Moreover, the user can export the comparison results to a fresh *Excel* sheet, he or she can access the dendrogram for the selected semantic field, view examples of the collocations in the examples pane, and look up words or phrases in the *Merriam-Webster* online dictionary

and thesaurus (w[5]). The dendrograms are also available and allow the user to compare the results to the meaning relationships within the semantic fields.

5.5 Cluster Comparison

The tab Cluster Comparison allows the user to find out whether there is a relation between the semantic similarity and the collocation range of two items. If the hypothesis were confirmed that bases that are very close in meaning or even synonymous can be used with the same collocators, translators would have a very powerful strategy at hand that could help them handle collocations.

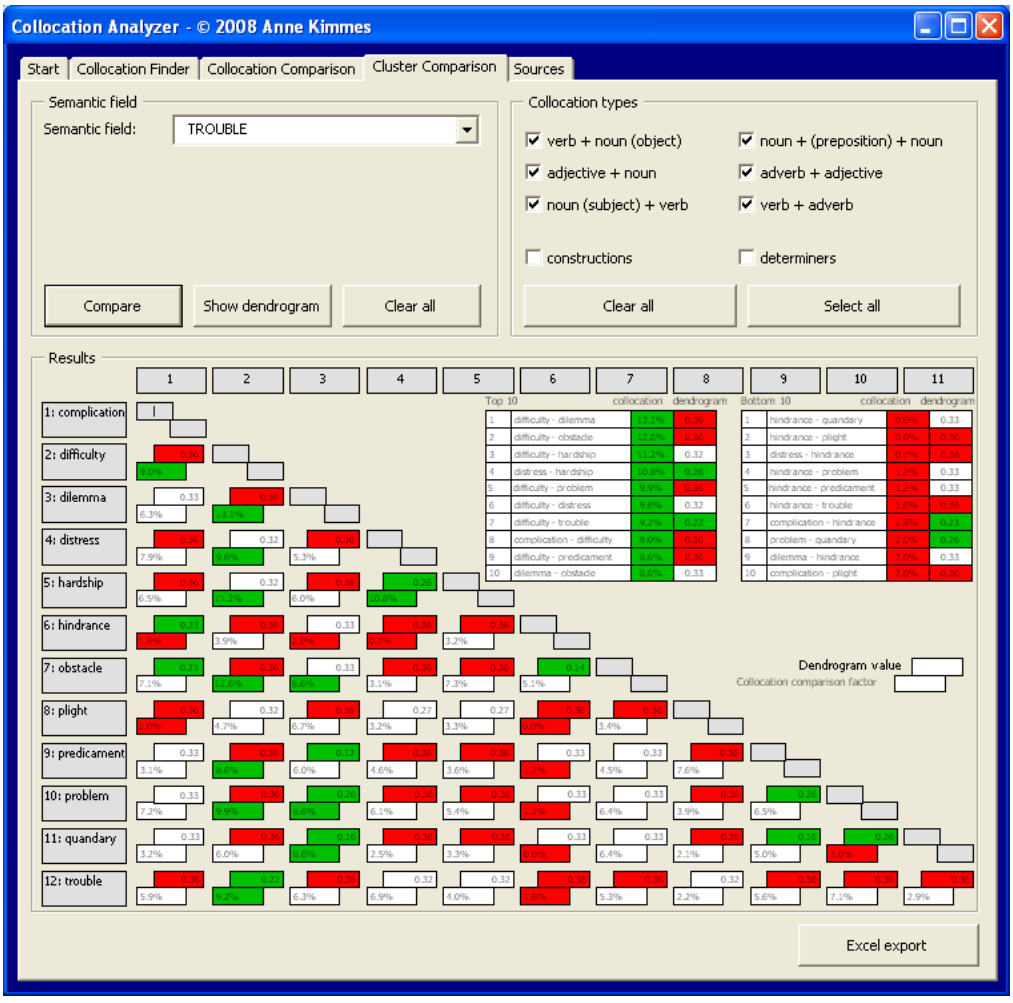

Fig. 8: Tab 4 of COLLOCATION | ANALYZER: Cluster Comparison

To test the hypothesis, the user must select a semantic field from the pull-down menu and click Compare. Once the data analysis is complete, a cross tabulation listing two values for all 66 pairs of bases within the chosen semantic field becomes available.

There are two data boxes for each pair of two bases. The value in the upper box indicates the value from the dendrogram. The value can be obtained from the scales available at the top and the bottom of the dendrogram (see Figure 6). The number in the box is the value at which the two items meet, i.e. where the two bases join a common cluster during the hierarchical cluster analysis. The smaller this value is, the more similar is the meaning of the two items.

The number in the box at the bottom indicates what percentage of collocations the two items have in common. If, for instance, one hundred different collocators were retrieved for the base "trouble" and also one hundred different collocators were retrieved for the base "complication" and the sources provided examples showing that twenty of those collocators could be used with both nouns, the percentage of common collocations would be ten percent (20 of 200).

In addition to the cross tabulation, the tab Cluster Comparison also offers two lists that contain the ten pairs of items that have the highest percentage of collocations in common (Top 10) and the ten pairs of items that have the lowest percentage of collocations in common (Bottom 10). The dendrogram value corresponding to the pair of bases is listed next to the percentages in the top ten and bottom ten lists.

All boxes in the Cluster Comparison tab contain an automatically generated color-coding. The ten lowest percentages and the ten highest dendrogram values are represented in red color. The ten highest percentages of common collocations and the ten lowest dendrogram values are colored in green. If two values are identical, they are all represented in the given color. For instance, it is very common that several pairs of items have the same dendrogram value because they all merge into a common cluster at the same point in the cluster analysis. It is important to remember that the higher the dendrogram value is, the smaller is the

degree of semantic similarity. Likewise the smaller the dendrogram value, the closer those two items are in meaning.

If the hypothesis were confirmed that bases that are very similar in meaning can be used with the same collocators, there should only be green boxes in the top ten list and only red boxes in the bottom ten list. Pairs with a high percentage of common collocations should have a high degree of semantic similarity (= a low dendrogram value) and pairs with no or very few common collocations should not be semantically related (= have a high dendrogram value). However, as can be seen, this hypothesis is unfortunately rejected. The pair with the highest percentage of common collocations has one of the lowest dendrogram values. Red and green coloring are present in both the top ten and the bottom ten lists.

The cross tabulation and the lists show that bases belonging to a semantic field do not have a corresponding field of collocators. There are only overlappings in their collocational ranges. Translators cannot follow the strategy of interchanging the collocators of semantically similar bases. This again proves that collocations are arbitrary units of a language and further emphasizes the urgent need for modern reference tools in this field.

As for the other options available in this tab, the cluster comparison can of course also be restricted to particular collocation types with the help of the checkboxes at the upper right-hand side of the interface, and the results can be exported to *Microsoft Excel*.

5.6 Sources

The last tab available within COLLOCATION│ANALYZER allows the user to define the sources from which he or she would like to obtain results. As has been said, the tool contains collocations from three different sources: the *Oxford Collocations Dictionary for Students of English*, the *Student's Dictionary of Collocations* and the *Washington Post* newspaper.

All of the collocations the two dictionaries contain for the 48 bases are available within the COLLOCATION│ANALYZER. However, the vast majority of the collocations available in the tool stem from the *Washington Post*. As explained above, a corpus consisting of all the news writing published on washingtonpost.com between March 1, 2006 and February 28, 2007

was searched for collocations of the 48 nouns. The search was performed either until the entire corpus (37,642,155 words) was scanned for collocations or until at least one hundred different word combinations were retrieved. The collocational data was transferred to an *Excel* database. Thus, COLLOCATION | ANALYZER does not directly query the two dictionaries or the newspaper corpus but the underlying *Excel* database.

Fig. 9: Tab 5 of COLLOCATION | ANALYZER: Sources

In the tab Sources, the user can select via the checkboxes if he or she wants to obtain collocations from one, two or all three of the sources. In the case of the *Washington Post*, the user can even choose to only access collocations from a particular part of the corpus

such as a specific month of publication or a particular newspaper section. Selections made in this tab directly influence the results of the other tabs.

6 Summary and Concluding Remarks

Collocations are a potentially large source of error for anyone who produces texts in a foreign language. Nevertheless, correct and idiomatic expression is indispensable for language professionals such as translators and interpreters. Since it is impossible to learn all collocations of a language (Angelone and Connelly 2007: 28, 31), translators need a powerful reference tool that helps them clear these linguistic hurdles. However, up to the present the reference works available are far from satisfactory. Collocations are inadequately listed in general dictionaries, and there are no specialized dictionaries in electronic format. The linguistic discipline is aware of this shortcoming and has submitted a number of proposals for a perfect reference work in the field of collocations. So far, however, none of them has been implemented and is ready to be used.

With COLLOCATION│ANALYZER, a user-friendly electronic reference tool in the field of collocations has been developed. It is a powerful tool that allows the user to perform various tasks: he or she can retrieve unknown collocations, verify tentatively available collocations, compare collocations of semantically similar bases, view contextual examples, sort and filter the results in a number of different ways as well as look up words in the renowned *Merriam-Webster* online dictionary and thesaurus directly from within the application.

The system requirements are minimal. The user simply needs a working version of *Microsoft Excel 2007* and to enable macros. Up to the present, COLLOCATION│ ANALYZER contains collocations of 48 noun bases belonging to four distinct semantic fields. There are approximately 20,000 entries. It is desirable to extend it to more central semantic fields or even to the entire vocabulary in the future. It is also conceivable to transfer the contents from *Excel* to a different database system such as an *SQL* or an *Access* database to handle larger amounts of data more easily. In much the same way as the contents of the COLLOCATION│ANALYZER can be extended, it is also possible to expand the functions. In the current version, there is a link to the *Merriam-Webster* online. It is imaginable to incorporate

other online reference sites into the tool in order to allow the reader to launch several online search queries directly from within the application. Besides monolingual and bilingual online dictionaries, COLLOCATION│ANALYZER could also link to collocation extraction sites such as *Cobuild Concordance and Collocations Sampler* or the *Leeds Collection of Internet Corpora*. COLLOCATION│ ANALYZER could also be extended to include one or several foreign languages.

Although there are still many possibilities to refine the reference tool, COLLOCATION│ANALYZER is already a powerful and fully functional application that will facilitate the translator's everyday work. It can assist the translator in making the appropriate selections in the syntagmatic dimension faster and optimize his choices. The corpus-based contents of the tool and its numerous contextual examples should dispel all doubts about inappropriate word combinations.

7 References

Angelone, Erik (2007). *The Conceptualization and Integration of an E-Collocation Trainer: Methods of Empirical, Translation-Based Collocation Research*. Joachim Kornelius and Jekatherina Lebedewa (eds.). *Heidelberger Studien zur Übersetzungswissenschaft* 7. Trier: WVT Wissenschaftlicher Verlag Trier.

Angelone, Erik and Martha Connelly (2007). "E-COR (Electronic Collocation Organization and Retrieval) as a CAT-Tool: A Computer-based Approach for Fostering English L2 Collocational Competence." Kerstin Brenner and Anja Holderbaum (eds.) in cooperation with Anne Kimmes. *Gebundener Sprachgebrauch in der Übersetzungswissenschaft. Festschrift für Joachim Kornelius zum 60. Geburtstag*. Joachim Kornelius and Anja Holderbaum (eds.). *Lighthouse Unlimited* 100. Trier: WVT Wissenschaftlicher Verlag Trier. 27-38.

Bahns, Jens (1996). *Kollokationen als lexikographisches Problem: Eine Analyse allgemeiner und spezieller Lernwörterbücher des Englischen*. Allén Sture et al. (eds.). *Lexicographica, Series Maior* 74. Tübingen: Niemeyer.

Benson, Morton, Evelyn Benson, and Robert Ilson (1997). *The BBI Combinatory Dictionary of English: A Guide to Word Combinations,* rev. ed. Amsterdam and Philadelphia: John Benjamins.

Benson, Morton, Evelyn Benson, and Robert Ilson (1999). *Student's Dictionary of Collocations*. Berlin: Cornelsen.

Butina-Koller, Ekaterina (2005). *Kollokationen im zweisprachigen Wörterbuch: Zur Behandlung lexikalischer Kollokationen in allgemeinsprachlichen Wörterbüchern des Sprachenpaares Französisch/Russisch*. Allén Sture et al. (eds.). *Lexicographica*, *Series Maior* 124. Tübingen: Niemeyer.

Crowther, Jonathan, Sheila Dignen, and Diana Lea (eds.) (2002). *Oxford Collocations Dictionary for Students of English*. Oxford: Oxford University Press.

Diller, Hans-Jürgen and Joachim Kornelius (1978). *Linguistische Probleme der Übersetzung*. Wolf-Dietrich Bald, Herbert E. Brekle, and Wolfgang Kühlwein (eds.). *Anglistische Arbeitshefte* 19. Tübingen: Niemeyer.

Firth, John R. (1957). "Modes of Meaning." John R. Firth (ed.) *Papers in Linguistics 1934-1951*. London: Oxford UP. 190-215. (Article originally published in 1951).

Firth, John Rupert (1968). "A Synopsis of Linguistic Theory, 1930-1955." Frank Robert Palmer (ed.) (1968) *Selected Papers of J. R. Firth 1952-59*. London: Longman. 186-205. (Article originally published in 1957).

Hausmann, Franz Josef (1984). "Wortschatzlernen ist Kollokationslernen: Zum Lehren und Lernen französischer Wortverbindungen." *Praxis des neusprachlichen Unterrichts* 31. 395-406.

Hausmann, Franz Josef (1999). "Praktische Einführung in den Gebrauch des *Student's Dictionary of Collocations*." Morton Benson, Evelyn Benson, and Robert Ilson (eds.) *Student's Dictionary of Collocations*. Berlin: Cornelsen. iv-xv.

Holderbaum, Anja (2003). *Kollokationen als Problemgrößen der Sprachmittlung*. Joachim Kornelius and Anja Holderbaum (eds.). *Lighthouse Unlimited* 30. Trier: WVT Wissenschaftlicher Verlag Trier.

Holderbaum, Anja. *Kollokationsdatenbank deutsch-englisch*. http://www.lighthouse-unlimited.de/kollokationen/.

Jehle, Günter (2007). *The Advanced Foreign Learner's Mental Lexicon: Storage and Retrieval of Verb-Noun Collocations like 'to embezzle money.' Philologia – Sprachwissenschaftliche Forschungsergebnisse* 93. Hamburg: Dr. Kovač.

Kimmes, Anne (2009). *Exploring the Lexical Organization of English: Semantic Fields and their Collocational Ranges*. Joachim Kornelius and Jekatherina Lebedewa (eds.). *Heidelberger Studien zur Übersetzungswissenschaft* 11. Trier: WVT Wissenschaftlicher Verlag Trier.

Kornelius, Joachim (1995a). "Über das Kollokationspotential in einsprachigen Lern-wörterbüchern am Beispiel des *Longman Language Activators* und des *Collins COBUILD English Language Dictionary*: Vom Printwörterbuch zum elektronischen Spezial-wörterbuch." Manfred Beyer et al. (eds.). *Realities of Translating*. *anglistik & englischunterricht* 55/56. Heidelberg: Winter. 313-27.

Kornelius, Joachim (1995b). "Vom Printwörterbuch zum elektronischen Kollokationswörter-buch: Theoretische, methodische und praktische Überlegungen zur Erstellung eines Kollokationswörterbuchs." Fredric F. M. Dolezal et al. (eds.). *Lexicographica* 11. Tübingen: Niemeyer. 153-71.

Longman Dictionary of Contemporary English, Writing Assistant Edition (2005). CD-ROM. Pearson Education Limited.

Steinbügl, Birgit (2005). *Deutsch-englische Kollokationen: Erfassung im zweisprachigen Wörterbüchern und Grenzen der korpusbasierten Analyse.* Allén Sture et al. (eds.). *Lexicographica*, *Series Maior* 126. Tübingen: Niemeyer.

w[1]: *SDL TRADOS – Translation Memory, Terminology Management and Software Localization*. http://www.trados.com/

w[2]: *Skype.* http://www.skype.com/

w[3]: *British National Corpus.* http://www.natcorp.ox.ac.uk/

w[4]: *Washington Post*. http://www.washingtonpost.com.

w[5]: *Merriam-Webster – Dictionary and Thesaurus Merriam-Webster Online.* http://www.merriam-webster.com/

w[6]: *Cobuild Concordance and Collocations Sampler*. http://www.collins.co.uk/corpus/CorpusSearch.aspx.

w[7]: *Leeds Collection of Internet Corpora*. http://corpus.leeds.ac.uk/internet.html.

This is a revised version of the article first published in: F. Austermühl, J. Kornelius (eds.). *Learning Theories and Practice in Translation Studies*. Trier 2008.