# Case study: creating a Russian-English Bilingual Dictionary from Corpus Data

## 1. METHODS

Broadly speaking, I adopted a three step approach:

1. Create monolingual definitions and categorisations.
2. Attempt to find matches across languages.
3. Create entries.

As we shall see, this approach is at heart monolingual and has some serious drawbacks when it comes to creating a bilingual dictionary.

### 1.1 Monolingual Categorisation

Two different approaches have been trialled:

1. Concordance-driven (message-focused)
2. Collocation-driven (form-focused)

Data processing was generally carried out on the basis of Russian and British National Corpora (RNC and BNC respectively). They are similar in size (147,577,522 and 110,691,482 tokens respectively) and, broadly speaking, in their approach to sampling (a wide cross-section of written and spoken genres, from various media and aiming to reflect the language as a whole). Occasionally, supplementary searches were conducted using the internet corpora for the respective languages (which are substantially bigger but restricted to a single medium).

All searches were performed using the interfaces available through *corpus.leeds.ac.uk.* Brodaly speaking, I gave preference to the *query interface* for its speed; however, more complex queries were often easier to construct in the *Intellitext* interface (Wilson, et al, 2010). *TheSketchEngine* often produced something of an information overload and was difficult to access off campus. The Birmingham Young University interface (*BYU-BNC*) is not terribly intuitive, and I did not have the time to invest into learning how to use it. This is unfortunate, because it appears to afford better access to medium/genre information than the other interfaces.

### 1.1.1 The Concordance-driven Approach

This is loosely based on Sinclair's (2003) 7-step procedure. I used samples of 100 concordance lines and attempted to categorise them according to the meaning and syntactic situation of the node.

I found that looking for easily identifiable patterns by eye (as Sinclair seems to suggest) produced quite poor results. The procedure yielded some of the more obvious syntactic information: for example, all the discovery verbs are often qualified by a small set of words describing ease/difficulty, possibility/impossibility, success failure; many are associated with particular prepositional constructions; some favour passive over active voicing or vice versa; most can qualify a subordinate clause. However, the above procedure revealed little in the way of pragmatic information. There are several possible reasons for this:

1. I am simply not very well practiced at spotting the subtler patterns.

2. The lexical fields are often diffuse: there are categories of words with clear similarities in function and meaning (for example, scientific instruments or techniques) that individually occur only once or twice. Collectively they just do not look similar enough to “pop out” on screen.
3. The patterns are obscured by syntactic variety. This is especially so in the case of Russian where the deviations from a (de facto) default word order are both permissible and often stylistically desirable.

As a consequence, I adopted a much more meaning-based approach: reading the concordances line by line and attempting to classify them predominantly by their informational payload rather than form. The ultimate goal, however, was the same as in the case of Sinclair (2003): to produce a number of well defined pragmatic/syntactic categories that allow for a complete mapping of the word’s lexical situation.

Broadly speaking, I attempted to categorise concordance lines according to:

1. Semantic prosody;
2. Domain;
3. Denotative meaning (what kind of action is being described).

I found that independently tracking further registral information was not useful, because it tended to be a function of the above. The categories were refined as I progressed, with back-checks after every major alteration.

Within this approach, collocation searches (sometimes with further concordancing) were predominantly used to check the relative prominence of various phenomena and to look for patterns that might have been concealed by the small sample size. However, they were informed by considerations that came out of a fundamentally qualitative analysis of concordance lines.

This approach was found to have a number of significant drawbacks:

1. Very time-consuming. I found that, even after some practice, it typically took upwards of four hours to process a single word.
2. Poorly framed categorisation. The categories were constructed on an *ad hoc* basis, sometimes resulting in very many categories, sometimes in very few, with considerable variation in the nature of categories. As a result, perfectly workable standalone definitions often became close to useless in a translation context, because it was difficult to isolate patterns of correspondence.
3. Not terribly reliable. Assignment to categories is often a question of subjective opinion and can vary with ephemeral factors like my mood.
4. One is at the mercy of sampling algorithms. In the case of one word (*заметить*; number 414 on the Wiktionary frequency list for Russian) over 80% of my sample came from collections of jokes. On a different day and a different machine, the same word came with more of a lean towards detective stories. (Luckily, there was no major shift in meaning or registral factors.)
5. Small sample sizes can obscure significant phenomena. My initial sample for *заметить* contained just one adjectival participle (roughly representative of the overall frequency of this particular form for this word). However, on examining a larger sample (500 lines), I discovered that adjectival participles are used in quite a distinct way and generally belong to a more formal/official register than other forms – the use in a joke was highly atypical.

### 1.1.2 Collocation-driven Approach

This was, arguably, closer to the fine detail of Sinclair's approach, except for the more automated nature of the procedure.

The procedure was, in essence, just a chain of collocation searches, starting with the most general possible (one word to the right, one word to the left) and funnelling in towards progressively more specific phenomena. Concordancing was used at various points as a "real world" check on this essentially abstract approach and also to inform hypotheses regarding meaning.

This approach has some significant advantages:

1. It is more quantitative and, therefore, more objective.
2. It allows for much broader sampling. (I am not sure whether the tools use actually search through the entire corpus, but it is safe to assume that they work with more than 100 lines).
3. The information is collected and presented in a way that makes it is easier to spot a broader range of phenomena.
4. The information on isolated syntactic and pragmatic phenomena is extremely convenient when it comes to cross-language comparison.
5. Theoretically, most of this procedure can be automated!

There are, however, also two significant disadvantages:

1. It is, in practice, remarkably difficult to get a consistent quantitative handle on what exactly is significant (see discussion of statistical tools below). This is especially so when the word comes associated with a diffuse lexical field (as in the scientific instruments/techniques example earlier). Low frequency collocates often come in sizeable groups associated with a narrow experiential domain, but there is no easy way

to process this information without human involvement ("acquisition bottleneck"; Bordag 2005).

2. Information that comes out of this procedure is often a classic case of blind men groping at an elephant: it is patchy and disjointed. It is hard to construct something resembling a coherent overall picture and even harder to judge whether this picture is more or less complete. Consequently, sporadic concordance checks still had to be used as a way of assessing progress and adjusting category definitions. On the plus side, many of the searches returned just a handful of lines, and, in the case of longer lists, a comparatively cursory version of concordance analysis was generally sufficient.
3. Technical problems are not uncommon. There is substantial variation between different search interfaces, and some features do not always work as designed.

I looked at only two of the four statistical parameters offered by the interfaces I was working with:

1. Mutual information index (MI) is an intuitively appealing measure but I came to the conclusion that the numbers would require significant further processing to allow meaningful conclusions. On its own MI rewards sheer obscurity (Bordag 2005).
2. The T-score largely follows base frequency, and, in the end, I simply couldn't justify adding a level of complexity without any immediately obvious benefit.

Ultimately, I came to a conclusion that, for this project, there was little to be gained by getting into statistical analysis. The scope of the project was just six words. Base frequency and common sense were enough to get the job done.

Needless to say, if I had to process larger volumes of data, a statistical tool of some kind would have been necessary.

### 1.2 Interlingual Matching

I tried to adopt what was basically a two step process:

1. Try to find roughly matching categories.
2. Search for corresponding collocations to refine the matches.

Again, this approach is, in its mindset, monolingual. In practice, it resulted in a rather chaotic and inefficient “back-and-forth” workflow.

As was discussed earlier, there were serious problems with the framing of the categories, so it was difficult to conducts systematic comparison. This stemmed from a deeper problem: different languages draw lexical boundaries according to different criteria. The workflow I eventually ended up with was the reverse of the above:

1. Use collocation searches designed with some sort of connotative equivalence in mind to look for candidates with relatively well defined domains (e.g. academic science, medicine, policing, crime reporting, nature documentaries etc.) and/or tenor (essentially, different degrees of formality).
2. Write bilingual definitions that are partly informed by monolingual ones but fundamentally different in the underlying mentality – less about neat philosophical or psychological categorisations and more about finding concise ways to identify areas of semiotic overlap.

For some of the more obscure word senses, adequate equivalents could not always be found in the corpora and translations had to be provided based on introspection alone (as in the case of the somewhat unsatisfactory *identify* = *классифицировать* [classify]).

### 1.3 Entry Creation

The process leading up to the writing of the definitions was briefly discussed in the last section.

The examples were mostly sourced from the respective corpora. Many were paraphrased for convenience. Where no convenient example presented itself, I used either the internet corpora or, in one or two cases, Google. Obviously, had I been trying to construct a serious lexicographic resource, I would have used unaltered examples of natural language, but, since I was only after a translating aid, it seemed hard to justify the extra baggage.

For reasons of time, the dictionary remains very much non-exhaustive. Some words are given more complete descriptions than others.

## 2. RESULTS

Below I discuss only the more prominent senses of the words examined in this study. For the rarer senses, please refer to my dictionary. For reasons to do with flaws in my methodology, what follows are essentially two separate discussions from a monolingual perspective.

## 2.1 English

The most obvious variation across the three discovery verbs is the variation in semantic prosody (Stubbs 1995), which can be summed up as follows:

| | | |
|---|---|---|
| *detect* | => | negative |
| *identify* | => | neutral |
| *spot* | => | positive |

This does not apply across the board and there are variations across different senses of each word (for example, *identify* acquires a clear negative prosody when you talk about identifying a person), but on a "bird's eye view" scale, this is true. More interestingly though, there are some finer attitudinal differences.

*Detecting* something normally implies two things: 1) you have some idea of what the thing you have detected is; 2) further action is required or, at the very least, expected. Moreover, while the event of detection is itself momentary, the implication is that one was in a state of raised alert at the time leading up to it, and that the detection itself was something of an observational feat. To *detect* something, one must assume the mindset of a lookout on a castle tower: your job is to scan the horizon for feint signs of things that are likely to require further action. It is a mindset of effortful attentiveness and vigilance. Consequently, to detect something in the scientific/technical context is to make a difficult and significant observation, but usually not an entirely surprising one.

In the case of detecting people's mood or other social variables, the word is being used sarcastically. The overt meaning is that somebody is that you are being uncommonly perceptive in noticing something well hidden about the

person. The subtext, however, is that that the person is behaving in an unsubtle, “readable” way.

*Identifying* something, unlike *detecting*, usually implies not an instantaneous event but a process of finite duration (even if not very long). Moreover, it is a process that requires cognitive exertion from the agent. In *Aktionsart* terms (Steiner 2004), this word is aptly classed as an “accomplishment”. Not only is there a clear end point, there is a feeling of “closure” about having establishing the identity of a previously unidentified thing – one has resolved an unsatisfactory situation. The mindset is essentially activity-focused and dispassionate about the outcome, hence the neutral semantic prosody.

*Spotting*, like *detecting*, is an instantaneous event (except in the special cases where it refers to a hobby or an ability, as in *trainspotting*). However, the word has a rather opportunistic undertone. *Spotting* is an event of unexpected instantaneous recognition. It happens not as a result of effortful attentiveness but as a result of being pleasantly alert – something akin to Csikszentmihalyi’s “flow”. In this case, there is a clearly positive semantic prosody. What people *spot* is not so much danger as potentially useful things: the other people and things about them, valuable objects, new information about their environment. To stick with the vocational metaphor, if the person who *detects* is a lookout, and the person who *identifies* is a scientist, the person who *spots* is a scout or a forager.

**2.3 Russian**

*Заметить*, in its most frequent sense, is similar to *spot*, in that it deals with instances of what you might call surprise recognition. However, there is much

more of an interpersonal component to it. If a person *spotting* or *detecting* is interacting with his environment, and a person *detecting* is interacting with an object, the *«человек замечающий»* [the "noticing person"] interacts primarily with other people. Part of the reason this word seems to be exceptionally common in jokes is that it tends to point towards social situations. People notice (or fail to notice) things like other people's mood, important details about their appearance and behaviour, potential sexual partners – all things with considerable comedy potential.

There is a much less commonly encountered sense of *заметить* which is much closer to detect in that it deals with noticing something that is potentially or currently bad – the observation puts the agent in a state of raised alert and requires further action. What is interesting is that almost exclusively in this sense that it appears as a passive adjectival participle (*замечен/замеченный*). One possible reason for this is that the passive constructions where this particular form of the word tends to surface have a heavy, official feel to them. In the rare cases where it does appear in a joke, the humour has fundamentally different mechanics: the funny thing is not the situation itself but the implication of non-trivial threat that is later revealed to be false, as in:

> *Yesterday an [insert nation of choice] submarine was detected in Russian territorial waters. Immediate action was taken: we went and took their oars.*

*Опознать* is a narrower and decidedly macabre version of *identify*. It tends to surface in the context of crime reports (people identifying bodies, criminals, stolen goods, etc.). Interestingly, there is also a less common neutral sense that is almost identical to the English counterpart – to establish the identity or

to gain knowledge of something. Fundamentally, there is actually no difference between the meanings: both describe a process of methodically extracting information an essentially passive object. This is probably what gives the word such an uncanny edge when applied to bodies, victims, and criminals – it is the reduction of a human being from an agent capable of self-identification to a passive object. Plus, the very fact that the identification of a human being becomes an effortful activity implies a degree of damage, decay, or deception. The association of *опознать* with death, violence and illicit behaviour is so strong that the word itself carries a pronounced negative affect: hearing it automatically implies that something bad either has or will happen to somebody (much as the word *affect*, in its everyday sense, implies not just impact on something but adverse, negative impact).

*Обнаружить* is a curious hybrid of *spot* and *detect*. Like *spot*, *обнаружить* tends to describe a surprise discovery; however, like with *detect*, the discovery is usually something that puts the agent in a state of raised alert. Thus there is no clear English translation for this word. My strategy was to use the relatively neutral *discover* to get across the immediate denotative meaning of the word and provide a recommendation for the translator to use a "vertical" approach (Nord 1997), i.e. to provide connotative equivalence by transferring the affective load onto another part of the text.

The other sense of *обнаружить* is an almost exact equivalent of the more neutral, "scientific" sense of *detect* with many of the same negative connotations in a medical context.

Interestingly, about one third of the instances of *обнаружить* in my sample are sarcastic. As with the sarcastic use of *заметить* discussed earlier, the

humour stems from the contrast between the implied alert and the trivial resolution.

## 3. CONCLUSION

There is a “mismatch” in the attitudinal boundaries between the English and Russian cohorts examined. The split between the English words is so clearly fits boundaries between functional mindsets that it practically begs for a Stephen Pinker style evolutionary psychology interpretation. In Russian things are not so neat: the colouring of words is less vocational and more reliant interpersonal – more about the agent’s response to other people, both positive (*заметить*) and negative (*опознать*). Thus, exact equivalence is rarely possible, and some elements of affect must be treated as “vertical translation units” and transferred onto other parts of the text.

Words: 3005

Works Cited

Gülzow, Insa, and Natalia Gagarina. *Frequency Effects in Language Acquisition: Defining the Limits of Frequency as an Explanatory Concept*. Berlin: Mouton De Gruyter, 2007.

Nord, Christiane. *A Functional Typology of Translations*. Amsterdam: John Benjamins, 1997.

Sinclair, John. *Reading Concordances: An Introduction*. London: Pearson/Longman, 2003.

Steiner, Erich. *Translated Texts: Properties, Variants, Evaluations*. Frankfurt Am Main: P. Lang, 2004. Print.

Wilson, James, Anthony Hartley, Serge Sharoff, and Paul Stephenson. Advanced corpus solutions for humanities researchers. In *Proc Advanced Corpus Solutions, PACLIC 24*, pages 36–43, Tohoku University, November 2010.