ОПЫТ СОЗДАНИЯ ЛЕКСИКО-ТИПОЛОГИЧЕСКОЙ БАЗЫ ДАННЫХ (НА ПРИМЕРЕ СЕМАНТИЧЕСКОГО ПОЛЯ БОЛИ)

CONSTRUCTING A LEXICO-TYPOLOGICAL DATABASE (FOR A STUDY OF PAIN PREDICATES)

Kostyrkin A. V. (languages@bk.ru),
Panina A. S. (panina-anna@yandex.ru),
Institute of Oriental studies, RAS

Reznikova T. I. (tanja.reznikova@gmail.com)
All-Russian Institute of Scientific and Technical Information, RAS

Bonch-Osmolovskaia A. A. (abonch@gmail.com)
Higher School of Economics

We present a database developed for lexico-typological study of expressions of pain (demo version available at http://orientling.ru/bolit/). Its design implements the non-relational, NoSql approach, where data is organized into a flexible tree not limited in size and depth. Linguistic annotation is placed directly into the text of example sentences and their translations, so that in effect the database is structured as an annotated corpus.

This formalism gives much freedom to both the developers in their task of annotating examples, and users in their queries, since it allows them to vary the level of detail according to how much information is available or needed.

Linguistic annotation includes tags for syntactic roles, some syntactic constructions and their components (relative clauses, light verbs, formal subjects, parts of compound words), morphological information (tags for case, number, aspect etc), as well as semantic tags specific to the domain of pain (semantic roles and types of metaphoric shift).

Key words: lexical typology, lexical semantics, database, NoSql, pain

1. Introduction

A prominent and promising method in lexical typology lately has been comparing semantically coherent word classes, rather than meanings of individual words, across a number of languages. This is the primary goal of the Moscow lexico-typological group, from the Aquamotion project [Maisak, Rakhilina 2007, Lander Y. et al. 2010] describing verbs of swimming and floating, to the present research into the vocabulary of pain [Britsyn, Rakhilina et al. 2009].

Invaluable in such comparative studies, especially when many languages are involved, are databases. They offer, first, a way to record any discovered rules and generalizations, to organize data and present it in a readily accessible form. Secondly, a database itself can be made into an analytical instrument, as it allows to search for new correlations, verify hypotheses etc. The advantages of a database compared to a mere collection of raw data (even digitized) are obvious, and in direct proportion to its functionality.

Naturally databases have been employed in lexical studies for some time, yet none of the existing analogues suited the specific demands of our project.

WordNet [WordNet Lexical Database 2012], whose sheer scope makes it the flagship of lexical semantics representation, is ontologically rather than typologically oriented. With each language essentially independent from the others within its loose general framework, it is not suited for — nor, indeed, aimed at — the discovery of cross-linguistic phenomena. The same can be said about FrameNet [FrameNet Project 2012] and Lexicograph [Lexicograph Project 2012]. While each database provides an in-depth description for important areas of English and Russian lexicon respectively, neither was ever intended to encompass more than one language.

On the other hand, Anna A. Zalizniak's Catalogue of Semantic Shifts [Gruntov I. 2007, Zalizniak 2009] covers a number of languages and aims explicitly to serve typologically relevant observations, but the information it provides is limited, for the most part, to the source and target domains of the semantic shifts and does not detail any accompanying differences or similarities in argument realization and other syntactic behavior.

It is natural that a database's structure and content are determined by its purpose. However, even the single task of describing a relatively tight thematic class of predicates across several languages turns out to call for more than one database design. In fact, each thematic class appears to present a different set of requirements for its formal representation.

The Moscow lexico-typological group views a thematic class such as verbs of animal sound emission, motion in liquids, or pain in terms of an underlying system of oppositions that define relevant aspects of the situation, and, optionally, its relations to other domains. For aquamotion the main task was to find how this little domain itself is organized in each particular language, with such defining features as agency of the moving object (as in swimming vs. drifting) and directedness of motion (as in being carried by a current vs. floating in place). Verbs of animal sounds proved interesting first of all as a source of metaphoric shifts into other kinds of sound, speech and noise, and the database for the project's data was built around an elaborate classification of human, artificial and natural sound-emitters. Neither meta-structure could be used as is for pain predicates. For these we needed to accommodate secondary pain verbs recruited from non-algetic domains, and to fully capture this process we also had to pay more attention to the predicates' syntactic patterning. In other words, we had to design a database specifically suited to the project.

2. The domain of pain

What sets pain expressions apart from motion or sound is the way this domain, consistently across languages, is made up of a core of a few primary pain predicates and a greater number of secondary pain predicates.

Secondary predicates originate in one of the following domains: 1) combustion (e. g. English *my eyes **burn***) and related processes involving high temperature (e. g. German *glühen* 'glow with heat', Crimean Tatar *qajnamak* 'boil'); 2) deformation or destruction, in particular impact with sharp instruments such as blades or needles (e. g. Chinese *yāobù **cìtòng*** lit. 'the side pricks') and quasi-instruments such as claws or thorns (e. g. German *meine Augen **beißen*** lit. 'my eyes bite'); 3) motion, and deformation through motion (e. g. Ukrainian *nogi **krutit'*** lit. '[it] twists the legs'); 4) sound (e. g. Chinese *dùzi **jiào*** lit. 'the stomach screams') [Britsyn, Rakhilina et al. 2009].

One characteristic of a pain expression, therefore, is the type of metaphoric shift that created it, and we wanted to be able to keep examples of the original, non-metaphoric usage where applicable.

Syntactically primary pain verbs are statives, and as such tend to take a single argument interpreted as a theme or patient. Many of the secondary pain verbs, on the other hand, originate from transitive activities, and any changes in argument realization accompanying their shift into the domain of pain are important.

The basic situation of pain involves three participants: the experiencer, the affected body part, and a cause or stimulus, and being faced with a range of possibilities for syntactic realization of each of these arguments meant we needed to capture syntactic patterning in more or less detail, preferably including some morphological information. E.g., verbal aspect in languages such as Russian (imperf. *kolot'* vs. perf. *kol'nut'*) is involved in distinguishing continuous and momentary pain.

On the other hand, semantic classes are not really relevant to our task. For the body part and the experiencer they are trivial (the experiencer in our data is always human; while an animal experiencer is theoretically possible, the distinction of human vs. animal does not seem to bear any particular significance). For the predicates we decided not to subdivide the algetic domain into subdomains to represent specific kinds of pain as a medical phenomenon. Such classification would run a high risk of being arbitrary, since the experience of pain is highly subjective. Instead it is usually enough to know the stimulus and rely on our extralinguistic knowledge of the effect a given stimulus is likely to produce (e. g. the reaction of eyes to soap vs. bright light vs. strain). Accordingly we recognize the importance of collecting and classifying stimuli, while the most important feature of an algetic predicate is considered to be its domain of origin.

Our prospective database user, a semantic typologist, could need any of these types of data in any combination.

3. Annotation

One crucial decision that was made from the start was adopting the NoSql approach for database representation, opposite to the relational one. Within this approach linguistic data is structured not as a table with fixed number of fields but as a flexible tree not limited in size and depth.

Neither in retrieving nor in inputting data by this method does one have to struggle with a fixed set of obligatory features, some of which may be irrelevant to a given sentence. In this formalism providing meta-information is a matter of applying tags directly to the words that warrant them, without pulling the original sentence apart. In essence, the database is organized as an annotated corpus of sample sentences.

This is a novel method for computational studies of lexical typology, and it offers incomparably greater freedom in both in the content (e. g., without any predetermined formal requirements on patterning, any given example sentence need not have all, or any particular subset, of the basic participants — experiencer, body part, stimulus — realized in any particular way), and in the annotation, which easily accommodates subsequent corrections and additions.

This ease of managing meta-information proved a particularly significant benefit. The project gradually came to encompass the data from more then twenty differently structured languages such as Korean, Japanese, Spanish, French, German, Czech, Ukrainian, Serbo-Croatian, Hindi, Chinese and Crimean Tatar; but even without additional languages the mere accumulation of data often led to substantial changes in annotation.

For example, initially we recorded patterns of argument realization in what was hoped to be an exhaustive list. We would tag the predicate with a case frame from the list, such as *s. d. o.* for taking a subject, a direct object, and a dative object, and repeat the *s.*, *d*. and *o*. as its arguments' tags. There were similar tags for verbs taking oblique objects, possessives and clauses.

It was not long before ellipsis transpired as a strong factor. Was the verb still to be tagged *s. d.o* when no dative argument, or no argument except the subject, was present in the sentence, just to show the predicate's combinatory potential? Including only examples with a full set of participants seemed impractical, as well as contrary to actual usage.

This and some other considerations eventually led us to stop tagging the predicate with its argument realization pattern at all, and mark argument realization where it was happening, i. e. directly on the arguments, in as much syntactic and morphological detail as necessary.

Incidentally this revision of the syntactic annotation helped to deal with the need to make generalizations over transitive verbs in ergative languages (formerly tagged *a. o.*) and nominative languages (tagged *s. o.*). The ergative subject tag *a.* was abolished, and arguments got separate tags for syntactic role and for case.

The predicate, on the other hand, was newly marked for being the pain predicate, and for part of speech to account for nominals (*pains and aches*) and adnominals (*an aching tooth* etc).

Relative constructions (as in French *J'ai la gorge qui brûle* lit. *'I have a throat that burns'*), anaphoric constructions, impersonal constructions with a formal subject, and compounds (such as *heartburn*) each also required a set of specific tags for their constituents.

Another major change happened with the list of metaphors for secondary pain verbs. The list grew with the addition of new languages, e. g. the metaphor of light was added for the Serbo-Croatian verb *sevati,* lit. '*shine*', which is used to describe rheumatic aches. But apart from merely expanding the list of metaphors, at one point we introduced some, if limited, internal structure, such as allowing the universal tag for destruction *DESTR* to combine with the tag *SELF* for spontaneity (as in *My head is splitting*), or *INSTRUM* for instrumental action (as in *stabbing pain*).

The general principle behind the non-relational formalism is to include as much information as is available — for example, there are tags to mark a noun for oblique *and* a particular case such as ablative *and* gender *and* number, if we want, — but none of it is required; so that when some features are not applicable to the data (e. g. the sentence is in a language that has no grammatical gender or number) there is no need to tag them with default values, and when there is little information available the tags can also be minimal. Even the most essential tags for the participants of the algetic situation, their semantic roles — *exp* for experiencer, *bp* for body part, *s* for stimulus and *p* for the pain predicate itself, — are only present when the respective participants are present.

The same is true for the database user's side. Exactly the same mechanism serves to make a query as general or specific as needed, while freely combining semantic, syntactic and morphological information. Thus, a user can search for all sentences with a body part and an experiencer, or an experiencer realized as a subject, or as a subject in the nominative case only, or for a specific lexeme in any role, but only in the plural, etc — the flexibility is endless.

There is no need to resort to predefined constructs such as "transitive verb with a dative object". The user can easily describe it for themselves by requesting a combination of conditions — namely, that the pain predicate be a verb, with a direct object and a dative object present in the same sentence — and, furthermore, has an opportunity to relax (or tighten) these restrictions, which would not be possible if the syntactic pattern was a single feature.

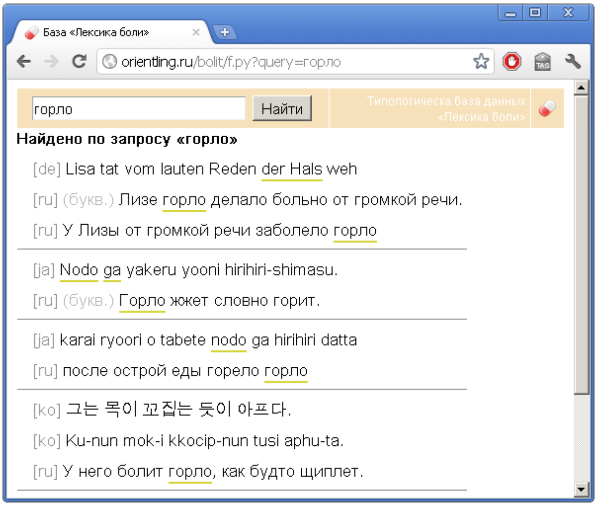
4. Implementation

Examples as they are appear in the search results (available as a working demo version at http://orientling.ru/bolit/) consist each of a sentence in national writing and/or transliteration, followed by its Russian translation, literal and/or literary:

The data as the annotator sees it when inputting and editing examples include the same plus the markup. E.g., the German sentence in Fig.1, *Lisa tat vom lauten Reden der Hals weh* ‘*Lisa’s throat hurt her from speaking loudly*’, looks as follows:

de: [[*Lisa*/exp/D/DAT=ru:*Лиза*]] [[*tat*/LV.DELAT/PAST.SG <*tun*=ru:*делать*]]
[[*vom lauten Reden*/st/OBL/ABL=ru:*громкая речь*]]
[[*der Hals*/bp/S/NOM.SG=1]] [[*weh*/p/ADV.GENERIC=ru:*больно*]]
ru: (букв.) *Лизе горло делало больно от громкой речи*.
ru: *У Лизы от громкой речи заболело* [[горло=1]]
#' (Jentzsch, K., Ankunft der Pandora)

Here we can see, enclosed in double square brackets and supplied with tags, the main unit of description: participant groups. The order of tags is not relevant, but traditionally we place semantic roles first. Beside the experiencer, stimulus, body part and pain predicate groups, there is one more group, not annotated with a semantic role — [[*tat*]], which is a purely syntactic part of the pain construction. The verb is marked *LV* for light verb, followed by its type *DELAT* ‘do’ (the other LV types are ‘be’ and ‘have’). The pain predicate, adverbial *weh*, is marked for *GENERIC* type, meaning non-metaphoric, primary pain word.


База «Лексика боли»
orientling.ru/bolit/f.py?query=горло
горло
Найти
Типологическа база данных
«Лексика боли»

Найдено по запросу «горло»

[de] Lisa tat vom lauten Reden der Hals weh

[ru] (букв.) Лизе горло делало больно от громкой речи.

[ru] У Лизы от громкой речи заболело горло

[ja] Nodo ga yakeru yooni hirihiri-shimasu.

[ru] (букв.) Горло жжет словно горит.

[ja] karai ryoori o tabete nodo ga hirihiri datta

[ru] после острой еды горело горло

[ko] 그는 목이 꼬집는 듯이 아프다.

[ko] Ku-nun mok-i kkocip-nun tusi aphu-ta.

[ru] У него болит горло, как будто щиплет.

Fig. 1. Search results for *‘throat’*

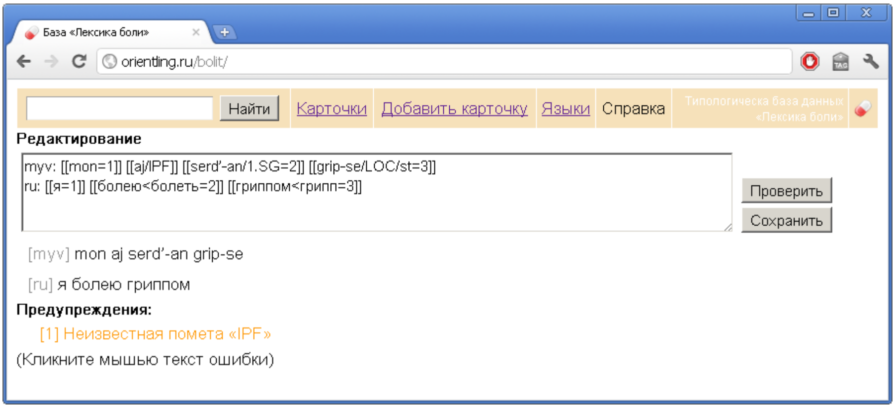
Head words in some of the groups are also annotated with the word’s dictionary form and Russian translation. This is done so that the sentence could be found, e. g., by searching for *tun* even though it only contains the inflected form *tat*. For *Hals* there is no need to place the Russian equivalent directly into the group in the source sentence, because, as Fig.1 proves, the example is searchable via the translation sentence.

The Russian equivalent is only of secondary importance, because the source lexical item is identifiable by its dictionary form.

Here also can be observed an experimental feature for better readability: *der Hals* in the source sentence and its equivalent *горло* in the translation are followed by an identifying number, which causes them to display underlined in a similar color and thus ties them together structurally and visually. The same notation can be used when one source word corresponds to several, not necessarily adjacent, words in the translation or vice versa. For example, to underline the correspondence of *tat* … *weh* lit. '*did* (her) *pain*' to *заболело* 'hurt' we just need to add another number to their respective groups:

de: [[*Lisa*/exp/D/DAT=ru:*Лиза*]] [[*tat*/LV.DELAT/PAST.SG <*tun*=ru:*делать***=2**]]
[[*vom lauten Reden*/st/OBL/ABL=ru:*громкая речь*]] [[*der Hals*/bp/S/NOM.SG=1]]
[[*weh*/p/ADV.GENERIC=ru:*больно***=2**]]
ru: *У Лизы от громкой речи* [[*заболело***=2**]] [[горло=1]]

There is a mechanism to check the notation for formal correctness when a sentence is added or edited. Besides errors which preclude an internal representation of the sentence from being formed (mostly syntactic, such as missing brackets), some situations which do not prevent the sentence from being parsed and saved into the database also cause a warning message to appear. One such message can be seen in Fig.2, warning about an unknown tag. In this case the tag is misspelled, but if it really was new and for some reason not yet on the list, it would not have prevented the example from being added.



Редактирование

myv: [[mon=1]] [[aj/IPF]] [[serd'-an/1.SG=2]] [[grip-se/LOC/st=3]]
ru: [[я=1]] [[болею<болеть=2]] [[гриппом<грипп=3]]

[myv] mon aj serd'-an grip-se

[ru] я болею гриппом

Предупреждения:

[1] Неизвестная помета «IPF»

(Кликните мышью текст ошибки)

Fig. 2. Edit window displaying a warning message for an unknown tag

To prevent tags from proliferating, annotators cannot add tags. Suggestions for any changes in the annotation are submitted to the database administrator; they are then discussed by all developers, and adopted as a joint decision.

5. Conclusion

In conclusion it can be said that the database of pain expressions, aside from its role as such, is also an attempt to build a flexible, versatile instrument for lexico-typological studies in general.

The purely grammatical notation, syntactic and morphological, is independent of the algetic domain and can be used elsewhere as is, or modified and expanded to accommodate more grammatical categories and syntactic constructions as new languages or tasks demands. The three semantic roles currently in use are, naturally, specific to the domain of pain and will have to be replaced with what is appropriate to the new frame describing the new semantic class. Similarly the list of metaphors, if a need arises to study words from the perspective of their origin in other domains, will need to be adjusted.

What we consider most significant, however, is the approach itself, embodied in our choice of non-relational formalism.

Not prescribing how a construction should be structured, not imposing predefined theoretical constructs onto data, but describing what is actually present in the text while allowing generalizations to naturally emerge as patterns of co-occurrence, it lets both the database compilers and users choose freely the amount of information to provide or demand, vague or specific, depending on their needs at the moment.

References

1. *Britsyn V. M., Rakhilina E. V., Reznikova T. I., Iavorskaia G. M.* eds. (2009) Concept of Pain in the Light of Typology [Kontsept bol' v tipologicheskom osveschenii]. Kiev.
2. *Gruntov I. A.* «The Catalogue of Semantic Shifts»: a Database for the Typology of Semantic Evolution. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2007" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2006"]. Bekasovo, 2007.
3. *FrameNet* Project (2012), available at https://framenet.icsi.berkeley.edu/fndrupal/
4. *Lander Y., Maisak T., Rakhilina E.* (2010) Domains of aqua-motion: a case study in lexical typology, in Motion Encoding in Language and Space, Oxford University Press, Oxford.
5. *Lexicograph* Project (2012), available at http://lexicograph.ruslang.ru/ (in Russian)
6. *Maisak T. A., Rakhilina E. V.* eds. (2007) Verbs of AQUA motion: lexical typology [Glagoly dviženija v vode: leksičeskaja tipologija], Indrik, Moscow.
7. *Zalizniak Anna A.* On the notion of semantic shift [O ponjatii semanticeskogo perehoda]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2009" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2009"]. Bekasovo, 2009, pp. 107–112.
8. *WordNet* Lexical Database (2012), available at http://wordnet.princeton.edu/