4

Language Documentation & Conservation Special Publication No. 16 (2019)
Methodological Tools for Linguistic Description and Typology
ed. by Aimée Lahaussois and Marine Vuillermet, pp. 45-61
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/24857

Automatic construction of lexical typological Questionnaires

Denis Paperno
Laboratoire Lorrain de Recherche en Informatique et ses Applications
(CNRS – Université de Lorraine – INRIA)

Daria Ryzhova
National Research University Higher School of Economics
(Moscow, Russia)

Questionnaires constitute a crucial tool in linguistic typology and language description. By nature, a Questionnaire is both an instrument and a result of typological work: its purpose is to help the study of a particular phenomenon cross-linguistically or in a particular language, but the creation of a Questionnaire is in turn based on the analysis of cross-linguistic data. We attempt to alleviate linguists' work by constructing lexical Questionnaires automatically prior to any manual analysis. A convenient Questionnaire format for revealing fine-grained semantic distinctions includes pairings of words with diagnostic contexts that trigger different lexicalizations across languages. Our method to construct this type of a Questionnaire relies on distributional vector representations of words and phrases which serve as input to a clustering algorithm. As an output, our system produces a compact prototype Questionnaire for cross-linguistic exploration of contextual equivalents of lexical items, with groups of three homogeneous contexts illustrating each usage. We provide examples of automatically generated Questionnaires based on 100 frequent adjectives of Russian, including *veselyj* 'funny', *ploxoj* 'bad', *dobryj* 'kind', *bystryj* 'quick', *ogromnyj* 'huge', *krasnyj* 'red', *byvšij* 'former' etc. Quantitative and qualitative evaluation of the Questionnaires confirms the viability of our method.

Keywords: lexical typology, adjectives, distributional semantic models, hierarchical clustering, questionnaire

E-ISSN 1934-5275

1. Introduction

Until recently, the lexicon was regarded to be an unsystematic and highly idiosyncratic part of a natural language, escaping any kind of a strict cross-linguistic comparison. This vision changed drastically when the seminal work on the typology of color terms by P. Kay and B. Berlin appeared in 1969. Since then, a growing body of research in so-called lexical typology (see Koptjevskaja-Tamm et al. 2016 for a recent overview) has consistently shown that a comparative study of words from different languages is a meaningful approach, but is fruitful only if there is a very well-defined *tertium comparationis* – a typological Questionnaire.[1]

Indeed, data extracted from different dictionaries are not in most cases directly comparable with each other, due to the lack of a single tradition and a unified metalanguage of dictionary entry representation. Monolingual and especially parallel corpora sometimes help to overcome the problem of data comparability (cf. Östling 2016; Wälchli & Cysouw 2012), but well-balanced corpora of a considerable size are only available for a very limited number of languages. In this situation, Questionnaires play a crucial role in typological research. Beside their main function, which is to provide uniform cross-linguistic data for a comparative study of a lexical domain, such Questionnaires can be also used in fieldwork as a tool for a typologically-oriented description of the lexicon in understudied, and especially endangered, languages.

There are several types of Questionnaires used in lexical typological studies: wordlists, checklists, translation-based questionnaires, sets of extralinguistic stimuli (pictures, video and audio clips, etc.). Construction of a Questionnaire of any of these types is time-consuming. Consequently, Questionnaires, especially those built with the goal of revealing fine-grained semantic distinctions, are usually designed for a very limited semantic domain, such as verbs of cutting and breaking (Majid & Bowerman 2007) or adjectives of speed (Plungian & Rakhilina 2013). To compensate for this and to describe the vocabulary of a low-resourced language, one needs a whole range of Questionnaires covering at least the core part of the lexicon.

In the present paper, we suggest a methodology to construct analytical questionnaires (which could also serve as translational ones; we elaborate on their possible uses in Section 2) for typologically-oriented lexicographic studies of words denoting qualitative features (corresponding to qualitative adjectives in English: *sharp*, *wet*, *warm*, and so on). Our technique is based on computational processing of a monolingual corpus. On the one hand, this method is fully automatic, and hence makes it possible to produce many

[1] We adopt here the terminology of the TULQuest project (cf. Lahaussois (2019) in this volume) and use the term "Questionnaire" (with the capital Q) to refer to any kind of written-based or extralinguistic stimuli used to collect linguistic data.

questionnaires very quickly. On the other hand, it is grounded in our experience of manual typological research, and we have tested our model on manually collected data from several semantic domains.

The paper is organized as follows. In Section 2, we present a brief overview of the existing types of lexical typological Questionnaires, analyse their advantages and shortcomings, and discuss in detail special Questionnaires that we seek to create automatically. In Section 3, we introduce the distributional semantic modeling framework, and in Section 4 we describe the results of our preliminary experiments of its application to the task of Questionnaire construction. Section 5 gives an overview of the 100 Questionnaires for adjectival lexicon produced automatically following the proposed methodology. Concluding discussion follows in Section 6.

2. Lexical Questionnaires

The most natural Questionnaires for lexical data collection, especially in the primary language documentation scenario, are wordlists of various kinds: the Swadesh list of core vocabulary and its versions adapted to specific regions (cf. Abbi 2001 for South Asia, Sutton & Walsh 1987 for Australia), the Intercontinental Dictionary Series (IDS) wordlist (Key & Comrie 2007), and others.

Despite the fact that the wordlists are primarily used to compile a dictionary for a particular language, data from different languages collected on the basis of one and the same set of concepts is used for typological studies as well. For example, the Database of Cross-Linguistic Colexifications (CLICS, List et al. 2014) is built primarily on data from the IDS, and the same list of concepts forms the basis of the World Loanword Database (WOLD, Haspelmath & Tadmor 2009).

Wordlists, however, are mostly oriented to nominal vocabulary. The IDS set of concepts is divided into twenty-four sections (kinship, animals, the body, the house, clothing and grooming, agriculture and vegetation, etc.), and only a few of them contain primarily verbal (motion) or adjectival (sense perception) notions. Words referring to concrete objects are easier to study and to elicit, because one can simply point a finger at their referents and ask a consultant to name every item, exactly as linguistic fieldwork guidelines recommend (cf. Bowern 2015). Differences within verbal and adjectival (or, more precisely, qualitative) semantic domains are much subtler and in most cases require additional typological research. As a result, the nominal lexicon is better elaborated and presented in more detail in lexical databases (cf. the very fine-grained representation of the domain 'earth – ground / soil – dust – mud' in the CLICS database), while the data on concepts that are usually expressed by verbs and adjectives is much poorer. Many such concepts are completely absent (e.g., 'swing' / 'sway' / 'oscillate'), and many others are too general (e.g. 'sharp', which in many languages is divided lexically into at least two subdomains: 'sharpness of cutting instruments' vs. 'sharpness of piercing instruments', cf. *tranchant* vs. *pointu* in French).

Typologists who focus on a particular lexical domain enjoy the opportunity to prepare a more detailed Questionnaire for the chosen semantic field. The best-known and most widespread lexical typological tradition is that of the research group at Max Planck Institute for Psycholinguistics in Nijmegen (Majid 2015). This approach is denotation-based: Questionnaires consist of carefully prepared extralinguistic stimuli of various kinds (pictures, video clips, sounds, etc.), and are hence easy to use in elicitation. A so-called "etic grid" forms the basis of every Questionnaire, i.e. sets of stimuli include all combinations of several parameter values. For example, the Munsell color chart is used to study color terms (cf. Berlin & Kay 1969; Kay et al. 2007), and video clips representing various combinations of subjects, objects (including also their possible final states) and instruments are designed for the analysis of verbs of cutting and breaking (Majid & Bowerman 2007).

This methodology allows for a very fine-grained analysis of certain semantic fields, with Questionnaires freely accessible and widely used in fieldwork. The main restriction of denotation-based Questionnaires concerns the range of lexical domains to which they can be applied: some concepts are hardly represented unambiguously with an extralinguistic stimulus, cf. evaluative meanings (good films, tasty food) or pain predicates. In order to take into account metaphorical extensions (cf. blue mood) and specific contextual constraints (e.g., the English colour term orange does not normally apply to hair color), additional techniques of data collection and analysis are required.

The frame approach to lexical typology, elaborated by the Moscow Lexical Typology group (Rakhilina & Reznikova 2016), relies on the linguistic behavior of the lexemes constituting a semantic domain and extends the Moscow Semantic School tradition of distinguishing between near-synonyms based on differences in their distribution (Apresjan 2000) to cross-linguistic comparison of translational equivalents. Within this methodology, groups of contexts referring to various types of extralinguistic situations ("frames") form a typological Questionnaire and serve as the tertium comparationis for the field in question. For example, the following situations are relevant for the domain 'sharp': 'sharpness of cutting instruments' (sharp knife, sharp blade), 'sharpness of piercing instruments' (sharp arrow, sharp spear), 'pointed form' (sharp / pointed nose, shoe toe), etc.

Frame-based Questionnaires are primarily analytical: they list possible usage patterns and predict potential lexical oppositions. They are intended for language experts who are supposed to fill them with language data from dictionaries, corpora and fieldwork, i.e. to find out what lexemes cover the semantic domain in question and what their contextual restrictions are. A list of minimal contexts that serve as illustrations for frames can be treated as a translation-based questionnaire, useful when working with bilingual consultants. However, it is recommended that language experts extend short diagnostic phrases to complete sentences or even paragraphs in order to provide a natural usage context example.

This methodology is applicable to any semantic domain, and allows for typological analysis of both direct and figurative senses of words. However, it comes at the price of the very time-consuming procedure of Questionnaire preparation. To reveal all the context types relevant to the field, one has to conduct a thorough investigation of contextual preferences of the lexemes from the chosen domain in at least 3-5 languages, based on dictionary and corpus data, as well as on native speaker judgments.

In the remainder of the paper we will present an algorithm inspired by the Frame approach procedure of Questionnaire construction that designs Questionnaires for words of qualitative features automatically. We highlight that our algorithm only uses data from one language (in our experiments we use Russian) as the input to typological predictions.

3. The approach taken: distributional models for semantic representations

Research in language typology suggests that typologically attested lexical distinctions are largely semantically motivated rather than idiosyncratic. If this is the case, one can find indications of potential semantic distinctions in any language, provided that different languages have comparable expressive power. One can therefore construct a lexical Questionnaire listing potentially distinct sets of word usages based on semantic representations for a single language. Such a monolingual Questionnaire might be approximate, in particular it may draw more potential distinctions than are attested in the lexica of actual natural languages (this can be treated as an additional bonus, unless these fine-grained oppositions are too numerous), or it may overlook some word usages that show a peculiar behavior in some languages. These potential drawbacks are compensated by the fact that such a Questionnaire can be built prior to any typological work.

One further advantage compared to the traditional typological research emerges if Questionnaires can be constructed automatically on the basis of computational semantic models. Here, we rely on distributional semantics (Lenci 2008).

The distributional approach to meaning represents each meaningful unit, typically a word, as a multidimensional vector (one can think of it as a point in a multidimensional space). The vector for each word is obtained from the statistics of the word's distribution in text corpora. While sharing these basic properties, distributional semantic models (DSM) come in different flavors that vary in their details. In some models, the dimensions of the vectors correspond directly to contexts (i.e. to the words that occur within a window of a certain size in relation to the target item), so that the value of a particular vector dimension is interpreted as a measure of association between the target word and the context. For instance, if dimension 537 corresponds to the context word *hand,* the value of dimension 537 for the vector of *bracelet* encodes the statistical association between the words *bracelet* and *hand.* More often, distributional models use latent vector representations from which one can predict the association between a word and its contexts but where the individual dimensions do not necessarily have such an immediate interpretation. Further, contexts can be collocates of a given word (e.g. *hand* as a context for *bracelet*) as in Lund & Burgess 1996, or documents in which the word appears, as in Landauer & Dumais 1997. A further dimension of variation within distributional models is the method used for obtaining the latent vector representations; this ranges from various analytical matrix decomposition methods, some of which are claimed to have greater interpretability than others (Griffiths et al. 2007), to neural models that learn semantic representations stochastically (e.g. Skip-gram and Continuous Bag of Words models, see Mikolov et al. 2013).

Distributional semantic models have shown good performance in various tasks and are generally known to contain a wealth of lexical semantic information. For example, a DSM can reliably predict human judgments about the semantic relatedness of words, and moreover it encodes cues about the properties of the words' referents (Herbelot & Vecchi 2015).

Distributional semantic representations have been extended beyond the meanings of words to larger meaningful units. In particular, multiple models for compositionality on word vectors and for their contextualization have been developed over the last decade. In the case of compositionality, the goal is to create a representation of a larger unit, such as the phrase *warm milk*, from representations of its parts, in this case from vectors of the words *warm* and *milk*. In the task of contextualization, on the other hand, the goal is to create a representation of a word meaning in a particular context, e.g. a representation of the meaning of *warm* when it occurs in the phrase *warm milk*. In practice, similar computational models have been applied to both tasks, with the simple vector addition serving as a good enough approximation of both compositionality and contextualization in many cases. There are mathematical reasons for the success of the additive model (Paperno & Baroni 2016), but its popularity derives mainly from the fact that increases in performance over addition come at the expense of considerably greater model complexity. This makes addition an obviously practical first choice for a compositionality model.

In our previous work[2] (Ryzhova et al. 2016) we proposed a new application of compositional distributional semantic models: predicting typological similarities between word usages, for example. We took usages of adjectives related to sharpness and smoothness as attested in the Moscow Database of Qualitative Features (Kyuseva et al. 2013). The Database stores data on lexicalization for approximately 20 semantic domains of physical qualities in an average of 15 languages. Language samples differ for different domains, but most of them include some Slavic, Germanic, Romance, Celtic and Finno-Ugric languages, as well as Mandarin Chinese, Japanese, Korean and some minor languages from the North Caucasus area. The Database contains a frame-based Questionnaire for every domain filled with data from the languages of the sample. Each usage type (*frame*) is represented by one or more diagnostic contexts which are nouns triggering a specific cross-linguistically invariant reading of the adjective when combining with it; for example, *nose* is one of the diagnostic contexts strongly associated with the 'pointed shape' reading of *sharp*.

Based on the data in the Moscow Database of Qualitative Features, we computed a measure of *typological closeness* for each pair of diagnostic contexts. Typological closeness ranges from 0 to 1 and characterizes the extent to which two contexts trigger the same lexicalizations of a property cross-linguistically. For example, the contexts *_knife* and *_blade* have a typological closeness of 1 for the property of sharpness, since sharpness of knives and blades is consistently lexicalized identically across languages. In comparison, while still falling within the same semantic field, in some languages there are distinct ways of expressing the sharpness of a stick and the pointed shape of a nose. Consequently, the typological closeness we estimated between the contexts *_stick* and *_nose* for the property of sharpness is only 0.82[3] (see an illustration in Table 1).

	English	French	Russian	Chinese	Besleney Kabardian (Circassian)
__knife	*sharp*	*tranchant*	*ostryj*	*fēnglì, kuài*	*ž'an*
__blade	*sharp*	*tranchant*	*ostryj*	*fēnglì, kuài*	*ž'an*
__stick	*sharp*	*pointu, aigu*	*ostryj*	*fēnglì, jiānlì, jiān*	*pamc̣e*
__nose	*pointed, sharp*	*pointu*	*ostryj*	*jiān*	*pamc̣e*

Table 1. Fragment of a Questionnaire for the domain 'sharp' filled with English, French, Russian, Chinese, and Besleney Kabardian data.

[2] An approach going in a similar direction is presented in (Koptjevskaja-Tamm & Sahlgren 2014).

[3] For the exact formula of typological closeness and other technical details see (Ryzhova et al. 2016).

Typological closeness of contexts was then compared against the vector similarity for compositional representations of corresponding Russian phrases, for example, the vectors for *ostraja palka* 'sharp stick' and *ostryj nos* 'sharp nose'. We rely on the standard measure of vector similarity, the cosine, which measures how close the directions in which the two vectors point are. We found that already the basic additive model of composition that simply adds the vectors *ostryj* 'sharp' and *palka* 'stick' to produce a representation of *ostraja palka* 'sharp stick' gives a high correlation between typological closeness and distributional vector similarity (65% Pearson correlation for the non-metaphorical usages of *sharp* and 74% for the non-metaphorical usages of *smooth*).

4. The algorithm

Our algorithm of Questionnaire construction rests upon two main assumptions. First, following the Frame approach to lexical typology, we believe that a Questionnaire for lexical typological research should contain types of contexts illustrating different types of word usage (frames). Second, based on the results of our previous research reported in Section 3 and additional experiments that we briefly review below, we assume that we can rely on the distributional semantic modeling technique to reveal the relevant context types automatically on the basis of a single language. We use Russian data in the experiments reported in this paper, but we assume that the language chosen should not affect the result in a major way.

We elaborated and tested the algorithm on typological data for several semantic domains of qualitative features ('sharp', 'straight', 'smooth', and 'thick') that were manually collected by experts from the Moscow Lexical Typology group (Luchina 2014; Kashkin & Vinogradova in print; Kozlov & Privizentseva in print; Kyuseva et al. in print). Previous research in the field demonstrates that the types of objects to which these qualities apply are in most cases responsible for cross-linguistic variation in the domains at hand. For example, a language can possess different lexical means to express the thickness of elongated vs. flat objects ('thick stick' vs. 'thick layer'), or to describe the age of human beings vs. artefacts ('old man' vs. 'old clothes'). Hence, the diagnostic contexts that form a Questionnaire are basic constructions consisting of the word denoting the qualitative feature and a nominal expression that it modifies. Similarly to English, qualitative features in Russian are usually expressed with adjectives, and the basic constructions take the form of noun phrases of the type "adjective + noun", with the adjective usually preceding the noun.

The algorithm that we propose to automatically design Questionnaires for qualitative features takes a list of Russian adjectives as an input and runs separately for every adjective from the list. The algorithm comprises the following steps:

1. collecting a set of nouns appearing no less than ten times next to the adjective in question in the main subcorpus of the Russian National Corpus (RNC, https://ruscorpora.ru);
2. computing a vector representation for every noun phrase ("adjective + noun" from the list collected at the previous stage);
3. clustering the distributional space of noun phrase vectors;
4. extracting three core elements from every cluster and eliminating all groups containing fewer than three elements.

Because Russian has a rich inflectional morphology, we use lemmas instead of word forms to collect a list of nouns and to compute all vector representations. We compose vectors for noun phrases from the co-occurrence vectors for every constituent using the simple additive composition model (Mitchell & Lapata 2010). To compute co-occurrence vectors, we count the occurrences of the 10 000 most frequent (according to the Russian National Corpus (RNC main subcorpus)) content words near the target lemma (within the context window of ±5 content words) in the RNC. To these raw co-occurrence vectors we apply the positive pointwise mutual information weighting scheme and reduce the dimensionality of the vector space from 10 000 to 300 dimensions using the singular value decomposition technique. We cluster the resulting distributional space with the hierarchical clustering algorithm that determines the optimal number of clusters automatically. To extract the core elements, we compute an average vector for every cluster and choose three noun phrases whose vector representations are the closest to the class centroid according to the cosine similarity metric.

To evaluate the resulting Questionnaires, we manually marked up the lists of noun phrases collected in the first stage of the algorithm's performance. For every noun phrase we indicated the frame it represented and then computed precision and recall for automatically designed Questionnaires. The recall values range from 0.733 to 1 for different semantic fields, implying that the Questionnaires included the vast majority of context types relevant for a domain. The precision values were in the 0.675 - 0.884 interval, showing that the clusters were quite homogeneous. See Ryzhova & Paperno (in print) for more details.

Since the results of the quantitative evaluation are sufficiently high, we assume that this methodology can help to create typological Questionnaires for a wide range of lexical domains, and the resulting Questionnaires should be more fine-grained and more useful for cross-linguistic semantic comparison of lexical data than existing wordlists.

5. Resulting Questionnaires: overview and error analysis

5.1 Overview

Following the method outlined above, we produced Questionnaires of adjective meanings based on the 100 most frequent adjectives in Russian. Adjective frequencies used for selection were taken from the fiction subcorpus[4] of the Russian National Corpus, as reported in (Lyashevskaya & Sharov 2009).

In the Questionnaires released with this paper, we list classes of usages for each input adjective of Russian. Each class is assigned an arbitrary identifying number and illustrated by three phrases along with the phrases' typicality scores for the given class. The typicality scores were computed as cosine similarity between the phrase vector and the class centroid. Each phrase is represented by lemmatized forms of the adjective and the noun separated by an underscore symbol ("_"). Note that lemmatization breaks the expression of agreement since the dictionary form of all adjectives is masculine. For example, the Questionnaire for *sčastlivyj* 'happy' contains a lemmatized entry *sčastlivyj_vstreča* 'happy encounter' with the adjective in a masculine form; of course, any natural texts will only use patterns with full agreement, such as *sčastlivaja* [nominative singular feminine] *vstreča*.

To make the set of Questionnaires easier to use, we divided the adjectives into several classes, using an automated clustering of adjective vectors followed by manual adjustments. The classes include adjectives of age, size, color, comparison, direction, location properties, order, social value, personality, emotional value, time, speed, temperature, and weight. Within the adjectives of size, we additionally group those that correspond to specific dimensions: depth, height, length, and width, as opposed to those qualifying size in general (e.g. *bol'šoj* 'big'). There are also 13 adjectives that do not fit well into any of these natural classes and are classified as 'other'. We naturally find borderline cases where some usages of an adjective could be attributed to a different semantic class than most usages. The semantic classification is therefore not intended to bear an independent scientific value and is applied only for the ease of use of our automatically constructed Questionnaires, as adjectives with related meanings are grouped in the same class.

Our work is intended to be evaluated where possible against reference data on lexical typology from the Moscow Database of Qualitative Features with Russian adjectives and context words used as keys. To enable an accurate comparison, we based our work on Russian corpora, and the Questionnaires contain Russian vocabulary in their entries. For

[4] Non-fiction is much less representative of everyday language usage than fiction. Word frequencies in non-fiction texts are skewed towards the official register, with *rossijskij* 'Russian' and *gosudarstvennyj* 'belonging to the state' being among the most common adjectives, making it to the top 15 list. Fiction gives a more natural frequency distribution.

illustrative purposes and to facilitate the adoption of our work for typological and lexicographic practice, we translated several of the Questionnaires (specifically, the *age* group) into English. The generated Questionnaires are available online:
→ https://sites.google.com/site/denispaperno/papers/questionnaires.zip.

5.2 Known errors

Both the selection of the adjectives and Questionnaire construction were carried out without manual intervention. Inevitably, the automatic procedure leads to some errors; for instance, the inclusion of adjectives 'Soviet' and 'Russian' in our list of frequent adjectives is an artifact of the reference corpus. The Questionnaires generated for them can nonetheless be useful for typological or lexicographic work. For example, one can think of *russkij* 'Russian' as a placeholder for the ethnonym adjective 'X' in language X. In this case different contexts in the Questionnaire for 'Russian' can be useful to reveal restrictions in usage of other ethnonym adjectives. To cite one distinction highlighted by the 'Russian' Questionnaire and relevant for the non-equivalent analogs of 'Russian', it is an enlightening semantic fact that an army or a fleet can be British but a language or a dress can only be English.

Some errors were introduced during Questionnaire construction, especially at preprocessing steps. We note multiple instances of incorrect lemmatization such as the missing ending in *duš* instead of the correct *duša* 'soul'. These cases should not constitute a major issue in practice since any linguist with knowledge of Russian will be able to immediately spot and correct them. We therefore warn future users about the existence of such glitches in our Questionnaires.

5.3 Initial qualitative analysis of the questionnaires

The Questionnaires created by our system tend to be quite fine-grained. For example, the Questionnaire produced for 'warm' (Russian *teplyj*) distinguishes 9 different classes of usages, presented in Table 2. These groups of contexts detect many situations that are known to be typologically distinct. The method captures two frames of temperature evaluation out of the three suggested in (Koptjevskaja-Tamm 2015): the TACTILE (water, food and drinks, body parts) vs. the AMBIENT temperature (weather objects, seasons, times of day). As for the third domain, that of the PERSONAL-FEELING temperature (cf. the English *I am hot*), it is quite expectedly absent from our list, because the related meanings cannot be expressed with an attributive construction in Russian. However, cluster 7 (clothes) relates to this frame in a metonymic fashion: applied to the nouns denoting clothes, the Russian *teplyj* means ‘helping to keep a comfortable PERSONAL-FEELING temperature when the ambient is cold’. In addition to the frames of temperature terms’ direct usages, the method captures their most common extended meanings, such as metaphorical social and emotional warmth, clusters 3 (‘warm company’, ‘kind (literally ‘warm’) concern’) and 9 (‘heart’ and ‘soul’) respectively. To compare, the Intercontinental Dictionary Series wordlist (Key & Comrie 2007) contains only one concept representing the whole ‘warm’ domain.

1: substances	*struja* ‘flow’	*vozdux* ‘air’	*voda* ‘water’
2: weather objects	*solnce* ‘sun’	*nebo* ‘sky’	*tuman* ‘mist’
3: social warmth	*kompanija* ‘company’	*no* ‘but’	*učastie* ‘concern’
4: food	*moloko* ‘milk’	*xleb* ‘bread’	*vodka* ‘vodka’
5: times of day	*utro* ‘morning’	*večer* ‘evening’	*noč’* ‘night’
6: seasons	*vesna* ‘spring’	*osen’* ‘autumn’	*zima* ‘winter’
7: clothes	*pal’to* ‘coat’	*kofta* ‘blouse’	*kurtka* ‘jacket’
8: human body parts	*ladon’* ‘hand’	*palec* ‘finger’	*plečo* ‘shoulder’
9: human body parts (metaphorical)	*serdce* ‘heart’	*duša* ‘soul’	*sleza* ‘tear’

Table 2. Example of a generated Questionnaire with usage classes for teplyj *‘warm’.*

The Questionnaires largely reflect the taxonomy of objects that the adjectives can describe, and for adjective meanings that have been studied cross-linguistically, our Questionnaires do make typologically attested distinctions between usages. To give one more example, the Questionnaire for *tolstyj* ‘thick’ differentiates the thickness of flat objects, long objects and fat humans, and these distinctions are indeed typologically relevant (Kozlov & Privizentseva in print).

We note that, somewhat surprisingly, a meaningful Questionnaire was also constructed for the adjective *nužnyj* ‘necessary’. This was not expected because *nužnyj* is predominantly used predicatively and differs in its distribution from most adjectives. Still, the algorithm managed to separate usages that seem to trigger different translation equivalents of *nužnyj* in English, distinguishing among others between usages such as *nužnaja vera* ‘**(much-) needed** faith’, *nužnyj dokument* ‘**required** document’, *nužnaja minuta* ‘**right** minute’, and *nužnoe dokazatel’stvo* ‘**necessary** evidence’.

Sometimes the classification of the adjective usages in a Questionnaire is too fine-grained and contains more classes than can be reasonably expected to show cross-linguistic differences in lexicalization. The extreme example here is ‘new’ (Russian *novyj*), for which our algorithm predicted 91 different classes of usage; among other things, the generated Questionnaire can be interpreted as distinguishing between novel church officials, new bosses, new hired managers, new monarchs, newly appointed military officials, and new judicial officials as all potentially requiring different lexicalizations of novelty. There is at least some truth to these hypothesized distinctions, as suggested by the existence of specialized adjectives such as the Russian *novopomazannyj* 'newly crowned', *novorukopoložennyj* 'newly ordained' and *novonaznačennyj* 'newly appointed' - all morphologically complex but lexicalized. We note however that the case of *novyj* is unique, probably due to the high frequency and extremely general semantics of 'new', and that all other adjectives have considerably fewer classes in our Questionnaires. An average Questionnaire contains 33 classes and the median number of classes in a Questionnaire is only 13. Three quarters of our Questionnaires include 20 or fewer classes of usages. For three adjectives, *poxožij* 'similar', *pozdnij* 'late', and *uverennyj* 'sure' the algorithm managed to identify only one class of usages.

6. Conclusion

We have presented an account of 100 automatically generated lexical Questionnaires for studying diverse usages of common quality-denoting vocabulary, as expressed by adjectives in English, Russian, and similar languages.

We believe that the Questionnaires presented will be immediately useful for linguists working on lexical data. In lexicography, they can provide input for typologically oriented dictionaries whose creation is of special importance for low-resourced and endangered languages. In turn, such dictionaries could become a basis for extensive cross-linguistic research in the future. For lexical typologists, our Questionnaires could be an insightful starting point saving much time and effort that are typically spent in the process of Questionnaire construction.

Of course, the resulting Questionnaires are not absolutely free from drawbacks. First, overly detailed clusterings could cause practical difficulties during fieldwork, as a Questionnaire for an interview with a consultant should be as short as possible. In future research, we plan to improve our method to reduce the number of context classes in overly long Questionnaires. Second, context-based Questionnaires require a non-trivial amount of work as they have to be translated into the languages studied. The translation process could also be automatized at least for languages with sufficient resources (see Ryzhova et al. 2018), but the translation algorithms we have tried so far require further improvements. Finally, our Questionnaires reflect some peculiarities of the Russian language and culture. From the lexical point of view, some nouns that appear in the final clusterings are culture specific (cf. *vodka* 'vodka' or *pal'to* 'coat' in Table 2), though the classes themselves are typologically relevant. From the syntactic point of view, the method in its current version is restricted to the meanings that can be expressed in Russian with an attributive construction. We will address both issues in our future research.

We hope that our automatically produced Questionnaires will be adopted by the linguistic community and will prove useful for lexical research of various kinds.

Acknowledgements

We thank the Editors of this volume, Marine Vuillermet and Aimée Lahaussois, and the anonymous Reviewers for their thoughtful and helpful comments on previous versions of this paper. The research has been supported by CNRS PEPS ReSeRVe grant (Denis Paperno) and by the Russian Foundation for Basic Research, grant 17-06-00184 A (Daria Ryzhova).

Appendix A. Adjectives used for the creation of Questionnaires

age: *molodoj* ‘young’, *staršij* ‘elder’, *novyj* ‘new’, *staryj* ‘old’

color: *belyj* ‘white’, *sinij* ‘blue’, *želtyj* ‘yellow’, *černyj* ‘black’, *seryj* ‘grey’, *zelenyj* ‘green’, *krasnyj* ‘red’, *temnyj* ‘dark’.

comparison: *ravnyj* ‘equal’, *raznyj* ‘different’, *poxožij* ‘similar’, *podobnyj* ‘analogous’

direction: *levyj* ‘left’, *pravyj* ‘right’

emotional evaluation: *čužoj* ‘foreign’, *krasivyj* ‘beautiful’, *prekrasnyj* ‘wonderful’, *dobryj* ‘kind’, *milyj* ‘nice’, *rodnoj* ‘native’, *dorogoj* ‘dear’, *nastojaščij* ‘real’, *sčastlivyj* ‘happy’, *xorošij* ‘good’, *nužnyj* ‘necessary’, *strašnyj* ‘horrible’, *jasnyj* ‘clear’, *ploxoj* ‘bad’

location properties: *blizkij* ‘close’, *dalekij* ‘faraway’, *tixij* ‘quiet’

order: *poslednij* ‘last’, *sledujuščij* ‘next’

personality: *spokojnyj* ‘calm’, *uverennyj* ‘confident’, *veselyj* ‘funny’

size: *bolʹšoj* ‘big’, *krupnyj* ‘large’, *nebolʹšoj* ‘small’, *ogromnyj* ‘huge’, *malenʹkij* ‘little’

  depth: *glubokij* ‘deep’, *melkij* ‘shallow’

  height: *nizkij* ‘low’, *vysokij* ‘high’

  length: *dlinnyj* ‘long’, *korotkij* ‘short’

  width: *širokij* ‘wide’, *tolstyj* ‘thick’, *tonkij* ‘thin’

social value: *čelovečeskij* ‘human’, *strannyj* ‘strange’, *detskij* ‘childish’, *osobyj* ‘special’, *svobodnyj* ‘free’, *glavnyj* ‘main’, *važnyj* ‘important’, *interesnyj* ‘interesting’, *prostoj* ‘simple’, *velikij* ‘great’, *izvestnyj* ‘well-known’, *russkij* ‘Russian’, *voennyj* ‘military’, *ser’eznyj* 'serious', *živoj* 'living', *lučšij* 'best', *sobstvennyj* 'own', *ženskij* 'feminine', *obščij* 'common', *sovetskij* 'Soviet'

speed: *bystryj* 'quick', *skoryj* 'fast'

temperature: *gorjačij* 'hot', *holodnyj* 'cold', *teplyj* 'warm'

time: *byvšij* 'former', *pozdnij* 'late', *nočnoj* 'happening at night', *rannij* 'early', *dolgij* 'long'

weight: *legkij* 'light', *tjaželyj* 'heavy'

other: *čistyj* 'clean', *polnyj* 'full', *celyj* 'whole', *golyj* 'naked', *pustoj* 'empty', *železnyj* 'iron', *gotovyj* 'ready', *znakomyj* 'familiar', *mokryj* 'wet', *edinstvennyj* 'only', *zolotoj* 'golden', *sil'nyj* 'strong', *obyčnyj* 'usual'.

References

Abbi, Anvita. 2001. *A manual of linguistic field work and structures of Indian languages*. Vol. 17. München: Lincom Europa.

Apresjan, Juri. 2000. *Systematic lexicography*. Translated from Russian by K. Windle. Oxford: Oxford University Press.

Berlin, Brent, & Kay, Paul. 1969. *Basic color terms: Their universality and evolution*. Berkeley: University of California Press.

Bowern, Claire. 2015. *Linguistic Fieldwork: A Practical Guide*. Basingstoke & New York: Palgrave Macmillan.

Griffiths, Thomas & Steyvers, Mark & Tenenbaum, Joshua. 2007. Topics in semantic representation. *Psychological Review*, 114(2). 211-244.

Haspelmath, Martin & Tadmor, Uri (eds.). 2009. *Loanwords in the World's Languages: A Comparative Handbook*. Berlin: Walter de Gruyter.

Herbelot, Aurélie & Vecchi, Eva Maria. 2015. Building a shared world: Mapping distributional to model-theoretic semantic spaces. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 22-32.

Kashkin, Egor & Vinogradova, Olga. (in print). The domain of surface texture. In Rakhilina, Ekaterina & Reznikova, Tatiana (eds.). *The Typology of Physical Qualities*. Amsterdam: John Benjamins Publishing Company.

Kay, Paul & Berlin, Brent & Maffi, Louisa & Merrifield, William R. & Cook, Richard. 2007. *World Color Survey*. Stanford: Center for the Study of Language and Information.

Key, Mary Ritchie & Comrie, Bernard. 2007. *Intercontinental dictionary series*. Online version: http://lingweb.eva.mpg.de/cgibin/ids/ids.pl.

Koptjevskaja-Tamm, M. (ed.). 2015. *The Linguistics of Temperature*. John Benjamins Publishing Company.

Koptjevskaja-Tamm, Maria & Rakhilina, Ekaterina & Vanhove, Martine. 2016. The semantics of lexical typology. In Riemer, Nick (ed.), *The Routledge Handbook of Semantics.* 434–454.

Koptjevskaja-Tamm, M., & Sahlgren, M. 2014. Temperature in the Word Space : Sense exploration of temperature expressions using word-space modeling. In Szmrecsanyi, Benedikt & Wälchli, Bernhard (eds.), *Linguistic variation in text and speech, within and across languages*. 231–267. Berlin/Boston: Walter de Gruyter.

Kozlov, Alexey & Privizentseva, Maria (in print). Typology of dimensions. In Rakhilina, Ekaterina & Reznikova, Tatiana (eds.). *The Typology of Physical Qualities*. Amsterdam: John Benjamins Publishing Company.

Kyuseva, Maria & Parina, Elena & Ryzhova, Daria. (in print). Methodology at work: semantic fields «sharp» and «blunt». In Rakhilina, Ekaterina & Reznikova, Tatiana (eds.). *The Typology of Physical Qualities*. Amsterdam: John Benjamins Publishing Company.

Kyuseva, Maria & Reznikova, Tatiana & Ryzhova, Daria. 2013. Tipologičeskaja baza dannyx adjektivnoj leksiki [A typologically oriented database of qualitative features]. In Selegei,V. P. & Belikov, V. I. & Boguslavskij, I. M. & Dobrov, B. V. & Dobrovolskij, D. O. & Zakharov, L. M. & Iomdin, L. L. & Kobozeva, I. M. & Kozerenko, E. B. & Krongauz, M. A. & Laufer, N. I. & Lukashevich, N. V. & McCarthy, D. & Nivre, J. & Osipov, G. S. & Raskin, V. & Segalovich, I. V. & Hovy, E. & Sharov, S. A. (eds.). *Kompjuternaya lingvistika i intellectual'nyye tehnologii* [*Computational Linguistics and Intellectual Technologies*]. Volume 1. 419–430. Moscow: Russian State University of Humanities.

Lahaussois, Aimée. 2019. The TULQuest linguistic questionnaire archive. In Lahaussois, Aimée & Vuillermet, Marine (eds.), *Methodological Tools for Linguistic Description and Typology*, Language Documentation & Conservation Special Publication No. 16. Honolulu: University of Hawai'i Press. 31-44.

Landauer, Thomas K. & Dumais, Susan T. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*. 104 (2). 211-240.

Lenci, Alessandro. 2008. Distributional semantics in linguistic and cognitive research. *Rivista di Linguistica* 20.1. 1-31.

List, Johann-Mattis & Mayer, Thomas & Terhalle, Anselm & Urban, Matthias. 2014. CLICS: Database of Cross-Linguistic Colexifications. Marburg: Forschungszentrum Deutscher Sprachatlas (Version 1.0, online available at http://CLICS.lingpy.org, accessed on 2018-6-23).

Luchina, Elena. 2014. Puti grammatikalizacii leksem so značeniem 'prjamoj' [Grammaticalization paths of lexemes with the meaning 'straight'] (Diploma paper). Lomonosov Moscow State University, Moscow.

Lund, Kevin & Burgess, Curt. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28 (2), 203–208.

Lyashevskaya, Olga & Sharov, Sergey. 2009. *The frequency dictionary of modern Russian language*. Moscow: Azbukovnik.

Majid, Asifa. 2015. Comparing lexicons cross-linguistically. In Taylor, John R. (Ed.), *The Oxford Handbook of the Word*. Oxford: Oxford University Press. 364–379. https://doi.org/10.1093/oxfordhb/9780199641604.013.020

Majid, Asifa & Bowerman, Melissa & Van Staden, Miriam & Boster, James S. 2007. The semantic categories of cutting and breaking events: A crosslinguistic perspective. *Cognitive Linguistics* 18, no. 2. 133-152.

Mitchell, Jeff & Lapata, Mirella. 2010. Composition in Distributional Models of Semantics. *Cognitive Science* 34(8). 1388–1429. https://doi.org/10.1111/j.1551-6709.2010.01106.x

Östling, Robert. 2016. Studying colexification through massively parallel corpora. In Koptjevskaja-Tamm, Maria & Juvonen, Paeivi (eds.). *The Lexical Typology of Semantic Shifts*. Berlin/Boston: Walter de Gruyter GmbH. 157–176.

Paperno, Denis & Baroni, Marco. 2016. When the whole is less than the sum of its parts: How composition affects PMI values in distributional semantic vectors. *Computational Linguistics* 42, no. 2. 345-350.

Plungian, Vladimir & Rakhilina, Ekaterina. 2013. Where do speed adjectives come from? *Russian Linguistics*, 37(3). 347–359. https://doi.org/10.1007/s11185-013-9117-7.

Rakhilina, Ekaterina & Reznikova, Tatiana. 2016. A frame-based methodology for lexical typology. In Koptjevskaja-Tamm, Maria & Juvonen, Paeivi (eds.). *The Lexical Typology of Semantic Shifts*. Berlin/Boston: Walter de Gruyter GmbH. 95–129.

Ryzhova, Daria & Kyuseva, Maria & Paperno, Denis. 2016. Typology of Adjectives Benchmark for Compositional Distributional Models. In *Proceedings of the Language Resources and Evaluation Conference*. Paris: European Language Resources Association (ELRA). 1253-1257

Ryzhova, Daria & Melnik, Anastasia & Yershov, Iliya & Panteleeva, Irina & Paperno, Denis & Singh, Yajuvendra & Sobolev, Mark. 2018. Automatic data collection in lexical typology. In *Proceedings of the international conference on Computational Linguistics and Artificial Intelligence "Dialog-2018"*. [Electronic publication: http://www.dialog-21.ru/media/4259/ryzhova_ershov_melnik.pdf]

Ryzhova, Daria & Paperno, Denis. (in print). Constructing typological questionnaire with distributional semantic models. In Rakhilina, Ekaterina & Reznikova, Tatiana (eds.). *The Typology of Physical Qualities*. Amsterdam: John Benjamins Publishing Company.

Sutton, Peter & Walsh, Michael. 1987. *Wordlist for Australian languages*. Canberra: Australian Institute of Aboriginal Studies.

Wälchli, Bernhard, & Cysouw, Michael. (2012). Lexical typology through similarity semantics : Toward a semantic map of motion verbs. *Linguistics* 50(3). 671–710.