The Linguistic Category Model in Polish (LCM-PL)

Aleksander Wawer and Justyna Sarzyńska
Institute of Computer Science, Polish Academy of Sciences, Jana Kazimierza 5, 01-248 Warszawa
Institute of Psychology, Polish Academy of Sciences, Jaracza 1, 00-378 Warszawa
Poland
axw@ipipan.waw.pl, jsarzynska@psych.pan.pl

Abstract

This article describes the first public release of Linguistic Category Model (LCM) dictionary for the Polish language (LCM-PL). It is used for verb categorization in terms of their abstractness and applied in many research scenarios, mostly in psychology. The dictionary consists of three distinctive parts: (1) sense-level manual annotation, (2) lexeme-level manual annotation, (3) lexeme-level automated annotation. The part (1) is of high quality yet the most expensive to obtain, therefore we complement it with options (2) and (3) to generate LCM labels for all verbs in Polish. Our dictionary is freely available for use and integrated with Słowosieć 3.0 (the Polish WordNet). Its quality will improve: we'll add more manually annotated senses and increase the quality of automated annotations.

Keywords: Linguistic Category Model, LCM, LCM-PL, Polish

1. Linguistic Category Model (LCM)

The LCM typology is a well-established tool to measure language abstraction, applicable for multiple problems as those listed in Section 1.1.. Its core idea is the categorization of verbs into classes reflecting their abstraction.

The most general, top level distinction of the Linguistic Category Model is the one between state verbs (SV) and action verbs. As LCM authors put it, *state verbs (SV) refer to mental and emotional states or changes therein. SVs refer to either a cognitive (to think, to understand, etc.) or an affective state (to hate, to admire, etc.)*. This verb category is the most abstract one and also present in Levin's typology.

The other more concrete type of verbs in the LCM are action verbs. This type is always instantiated as one of its two sub-types, descriptive and interpretative action verbs (DAV and IAV) that all refer to specific actions (e.g., to hit, to help, to gossip, etc.) with a clearly defined beginning and end. SVs,in contrast, represent enduring states that don't have a clearly defined beginning and end.

The distinction between DAVs and IAVs is based on double criteria. The first states that DAVs have at least one physically invariant feature (eg. to kick - leg, to kiss - mouth), whereas IAVs do not (therefore, are more abstract than DAVs). The second criterion, sentiment, states that IAVs have a pronounced evaluative component (e.g., positive IAVs such as to help, to encourage vs. negative IAVs such as to cheat, to bully), whereas DAVs do not (e.g., to phone, to talk). Descriptive action verbs(DAVs) are neutral in themselves (e.g. to push) but can gain an evaluative aspect dependent on the context (to push someone in front of a bus vs. to push someone away from an approaching bus). In practice, the criteria sometimes overlap. Some verbs have physical invariants but also have clear evaluative orientation. For instance, "to cry" always involves tears (an invariant physical feature), but carries negative sentiment.

1.1. Applications

Years of research has yielded considerable evidence that language abstraction is related to many psychological phenomena such as intergroup bias, stereotypes, expectancy bias or even human personality. LCM has proven to be an adequate tool for evaluating such phenomena. Wigboldus Semin and Spears (2000) (Wigboldus et al., 2000) used the LCM in their study which proved that describers use more abstract language for expectancy-consistent behaviors. This effect of expectancies is manifested in linguistic intergroup bias (LIB) wherein people encode and communicate desirable in-group and undesirable out-group behaviors more abstractly than undesirable in-group and desirable out-group behaviors (Maass et al., 1989). The same study found that abstract versus concrete communication play an important role in the perpetuation of stereotypes. In addition to expectancies, describers' goals such as self-presentational goals (Rubini and Sigall, 2002), the desire to compete or co-operate (de Montes et al., 2003) or willingness to protect one's group from threat (Maass et al., 1996) may affect one's level of language abstraction. The LCM was also used in studies demonstrating that language abstraction conveys information both about the person whose behavior is being described and also the describers themselves. Douglas and Sutton (2006) (Douglas and Sutton, 2006) asked participants to view a series of cartoons, each depicting a person performing a behavior deemed positive or negative as well as to reading a description of the behavior . They were then tested whether participants' judgments of describers' communication goals were affected by language abstraction. Participants were asked to rate the likelihood that the describer wanted to create a positive, negative and unbiased impression of the actor. The results show that describers who use relatively abstract language to describe others' behaviors are perceived to have biased attitudes and motives compared with describers who use more concrete language. Even personality differences may manifest themselves in language use. In a study conducted by Beukeboom, Tanis and Vermeulen (2012) (Beukeboom et al., 2013) participants' spontaneous verbal utterances in face-to face interactions were analyzed for language abstraction by applying the Linguistic Category Model. Results showed significant positive correlations between extraversion and language abstraction. The findings suggest that

the verbal style of extraverts is characterized by a higher level of abstract interpretation, whereas introverts tend to stick to concrete facts.

1.2. Previous Work in Polish

The first experiments with automated classification of Polish verbs according to the Linguistic Category Model were reported in (Rogozinska and Wawer, 2013). The research focused on a small set of 1170 verbs, translated from the English General Inquirer (Stone et al., 1966) `http://www.wjh.harvard.edu/~inquirer/` dictionary into Polish. The paper reported high level of agreement of LCM tags between English verbs and their Polish translations, as Kappa scores ranged from 0.83 to 0.87 depending on person (experiment involved two linguists).

However, translating verbs from the General Inquirer dictionary into Polish and copying their LCM labels, was not a satisfactory method to obtain a complete LCM dictionary for Polish due to poor coverage. As mentioned earlier, the 1170 verbs yielded by this method still missed about 90 percent of verbs present in the Polish WordNet (Słowosieć). Therefore, the goal of the experiments in (Rogozinska and Wawer, 2013) was to design and test automated methods of recognizing LCM classes from a sample of verb occurrences from the National Corpus of Polish. Authors applied machine learning to predict LCM class of a verb, based on various features designed using WordNet that explored mostly hyperonymy of nouns immediately following each verb, assuming such nouns as verb's arguments.

The results turned out promising in the sense of exceeding baselines, but due to low precision in recognizing DAV verbs (depending on the setup, precision was at 0.38 and 0.5) the method could not be assumed as satisfactory for use as the main source for LCM labels for the final dictionary.

2. The Polish LCM (LCM-PL): Methodology

2.1. Introduction

The efforts described in this paper are aimed to create the LCM dictionary with following goals: high quality, wide coverage and integration with the Polish WordNet. In this section we discuss these points in more detail.

2.2. Ensuring Quality

. We put significant efforts to achieve high quality LCM labeling. In particular, this rules out purely automatic verb labeling, at least when applied as reported in (Rogozinska and Wawer, 2013) due to low precision.

Ideally, we would like all our LCM labeling to be human-made, possibly by two independent annotators and a third (gold) one for resolving conflicts. However, due to budgetary constraints, this is currently not feasible and we manually annotated only a subset of Polish verbs, selecting the most frequently used verb senses. Also, at the moment our human annotations are single, not double.

2.3. Wide Coverage

. Contrary to previously described experimental works in Polish, we aimed for an LCM dictionary including as many word forms as possible, providing high quality for those most heavily used.

We used the frequency list of word lemmas (base forms) obtained from `http://nlp.pwr.wroc.pl/en/tools-and-resources/resources/frequency-list`. Using that list, we assigned frequencies to verb list found in Słowosieć 3.0.

2.4. Integration with Słowosieć (the Polish WordNet)

. For multiple reasons we decided to integrate our dictionary with the popular existing electronic dictionary (Piasecki et al., 2009). The most recent version 3.0 contains 32448 lexical entries for the verb part-of-speech ("czasownik").

Allowing LCM labels to be attached to WordNet's senses opens up interesting possibilities for high precision LCM labeling. From our study in Section 3. it follows that multiple senses of the same verb might, and indeed often have, different LCM tags. This is also the case for English language LCM annotation in the General Inquirer Dictionary (Stone et al., 1966). Therefore, aiming for precision requires providing sense-level rather than lexeme-level LCM tags.

Attaching LCM tags to WordNet senses is a difficult task. Słowosieć (The Polish WordNet) suffers from notoriously high number of senses, difficult to distinguish even for a professional lexicographer. It's often not easy to image the actual uses of some of the more rare senses.

Fortunately, the creators of Słowosieć provide a tool for automated word sense disambiguation. It takes Polish language texts as input and returns their senses as they are recognized in actual sentences (Kędzia et al., 2015).

Currently, our LCM dictionary contains manual annotation for 10000 verb senses, taken from the latest Słowosieć 3.0. We plan to gradually increase this number, prioritizing verb annotation by their frequency.

Unfortunately, domain of verbs (a property of verbs annotated in Słowosieć 3.0) is not usable for LCM. It seems that „cst" category consist only of verbs connected with enduring state or change of state and execlude verbs which refer to emotional or mental states of a subjective nature – the core of SV verbs according to the LCM Manual.

2.5. Annotation Guidelines

In the spirit of the English language annotation in the General Inquirer, where LCM tags were attached to word senses rather than words (lexical entries), we followed the same approach. Our sense inventory for verbs is based on Słowosieć 3.0 (Piasecki et al., 2009).

We provided annotators with the most recent, unpublished version of the LCM annotator guidelines (Schmid et al., 2017), obtained directly from prof. Klaus Fiedler, the author of LCM (Semin and Fiedler, 1988).

To annotate verb senses with LCM tags, we presented annotators all WordNet synsets of each verb. For each synset, we listed glosses and all synonyms (lexical entries) of that synset so that annotators could distinguish their meaning. We asked to annotate each sense with an appropriate LCM tag.

LCM	lexical id	synset id	domain	synsets / gloss
SV	81612	56818	state	sharing
DAV	89828	63657	ownership	share sth.
IAV	89829	63654	social life	divide, separate
DAV	89826	63655	change	divide
DAV	89827	56841	ownership	separate
DAV	81339	56584	thinking	Determine the quotient of two numbers

Table 1: LCM tags for the senses of Polish verb *dzielić* (eng. to divide)

LCM	lexical id	synset id	domain	synsets / gloss
IAV	85002	69644	social life	get lost
IAV	85000	55371	thinking	disappear, blur
IAV	85001	59688	ownership	fade away
SV	84630	69641	state	fade away
IAV	11550	4361	competition	die, lose life
DAV	22108	64281	change	break-up, perish, lose

Table 2: LCM tags for the senses of Polish verb *zginąć* (eng. to die)

2.6. Annotator Agreement in LCM

The lexeme-level agreement between annotators was reported in (Rogozinska and Wawer, 2013). The authors measured Cohen Kappa agreement between two Polish annotators, and also between LCM labels by each of the annotators and LCM labels of English equivalents of Polish verbs. The experiment has been performed on 1170 verbs taken from the General Inquirer `http://www.wjh.harvard.edu/~inquirer/` dictionary and translated into Polish. It adresses not only the issue of differences between two persons annotating the same verbs in one language, but also the difference between LCM labels of the same verb, translated into another language.

The results indicated that the Kappa between two Polish annotators was at 0.78, while the Kappa between their Polish LCM verb annotations and English LCM equivalents was between 0.83 and 0.87. As (Rogozinska and Wawer, 2013) concluded, the agreement is reasonably high, but it is also clear that the task is far from entirely easy and free from ambiguities.

In our paper we did not compute sense-level agreement, but we assume that it is likely that the level of agreement may be a bit lower than the one for lexeme-level annotations due to difficulties in understanding fine-grained senses and shades of meaning.

3. LCM and word senses

This section contains the description of selected verb senses, annotated with LCM tags. To illustrate sense-level annotations for LCM, let us focus on two verbs, picked from our dictionary: *dzielić* (eng. divide or share) in Table 1 and *zginąć* (eng. die) in Table 2. In each of the tables, column called 'domain' contains Polish WordNet verb domain of a specific sense `https://en.wikipedia.org/wiki/PlWordNet`.

Let our example discussion be based on the verb *dzielić* ('to divide'or 'to share') which has multiple senses that illustrate its various meanings spanning across all possible LCM labels.

The most abstract sense has the LCM label SV and its corresponding WordNet domain is state. It's English equivalent is 'to share'. In Polish, it refers to an abstract property of something being shared between multiple objects or people. For instance, in programming, an object reference may be shared between multiple class instances. A point of view may be shared between multiple people. No physical correlates are involved and the meaning is clearly an abstract one too, therefore SV label is the most appropriate.

In its sense related to social life domain, the verb becomes interpretative (IAV). It's English equivalent in this case means 'to divide'. An example of meaning reflected here may refer to groups of people divided by their opposite opinions, often linked to strong sentiments. There are no physical correlates and no objects are involved, therefore IAV tag is appropriate.

Finally, the verb may give a description of an observable event in a situational context. For example, a separation of ownership (eg. ownership of something is divided between multiple owners). This situation usually refers to some owned entity, therefore in this meaning the verb becomes a DAV.

Generally, the principles behind LCM labels make the distinction between IAV and DAV sometimes vague. If a verb refers to observable events in a situational context, but requires additional interpretation and evaluation, it is an IAV. Otherwise, we assumed it's a DAV, especially if some physical correlates may be found. As for the verb 'to share', some of its meanings rely on context, whereas other meanings possess an autonomous, context-independent meaning.

4. The Polish LCM (LCM-PL): Current State

Currently, the dictionary is available both as one downloadable file as well as in three separate pieces that reflect its structure (manual sense-level, lexeme-level annotations and automated annotations). The most recent version of the dictionary and its components are maintained at `http://clip.ipipan.waw.pl/LCM-PL`.

LCM	part (1)	part (2)	part (3)
SV	12%	4%	0%
IAV	35%	20%	2%
DAV	52%	75%	98%
count	10000 verb senses	1200 verbs	9200 verbs

Table 3: State of Polish LCM dictionary (LCM-PL)

- Part (1) Contains sense-level manual LCM annotations.
- Part (2) Contains lexeme-level manual LCM annotations.
- Part (3) Contains lexeme-level automatic LCM annotations.

The most recent state of the dictionary is reflected in Table 3 that contains percentages of LCM tags and verb (also verb sense) frequencies for manually annotated LCM tags.
Part (2) originates from The General Inquirer verbs with LCM tags translated into Polish. The translations have been manually corrected to ensure that the Polish verbs match English versions and their LCM tags were adjusted for Polish.
Part (3) reflecting automated LCM labels are generated using word embeddings and a neural network. We intend to provide the most recent evaluation (in terms of accuracy) of this part of the dictionary at the URL of the resource at `http://clip.ipipan.waw.pl/LCM-PL`.
The current state of automated predictions is as follows. For predicting LCM tags in (3) we used word2vec word embeddings of size 300 as verb representations. Embeddings were trained on the National Corpus of Polish (`http://www.nkjp.pl`) and Polish Wikipedia. To automatically predict LCM tags, we applied a two-layer (each size 300) perceptron-style neural network with selu activations that takes embedding of a verb as its input and predicts its output LCM tag. We measured the accuracy of this solution using 10-fold cross-validation on the data set from (2). The network, after hyper-parameter space optimizations, reached the average accuracy of 91.07 with standard deviation between folds at (+/- 2.73).
Our neural network solution turns out to be significantly better than results obtained using machine learning and hand-crafted features reported in (Rogozinska and Wawer, 2013). The evaluation data set was the same in both cases: the list of 1170 manually-labelled, lexeme-level, verbs taken from the English from the General Inquirer dictionary and translated into Polish.
Our results obtained using this enlarged data set and neural network approach are on similar or even slightly better level than the agreement between human annotators reported in (Rogozinska and Wawer, 2013). Despite this, we plan to increase the quality of predictions even further by experimenting with other neural architectures such as convolutional networks or the newest regularization algorithms.
One interesting conclusion from Table 3 is decreasing percentage of SV verbs between parts 1, 2 and 3, and increasing percentage of DAV verbs. This reflects the intuition that there are relatively few state verbs but they are among the most frequently used words. And on the contrary, there are many descriptive verbs, infrequently used, closely related to nominal meaning (sharing material correlates). The amount of IAV verbs is also decreasing with the frequency of usage.

5. Conclusions

In this paper we described the initial release of Linguistic Category Model (LCM) dictionary for the Polish language (LCM-PL), we believe the first widely usable version of the resource intended to measure the level of language abstractness in Polish.
The previous research on this topic (Rogozinska and Wawer, 2013) (especially in automated LCM dictionary generation) demonstrated limited usability of machine learning to automatically obtain LCM labels. In our paper we introduced a resource that is annotated manually in its most important parts: verbs whose senses that are used most frequently. The remaining part of Polish verbs, those less frequently used, was a subject of annotation on the level of lemmas, manual and automated – using word embeddings and a neural network. The quality of automated annotations has been evaluated and is comparable to human-level annotations.
In the future, we plan to further extend the manually annotated part of the dictionary. We also plan to increase the size of the manually labelled sense-level dictionary and manually verify more lemma-level annotations. We estimate that the size of the manually annotated part will not have to be larger than twice the size of the current state of the dictionary (as of early 2018).
The reason behind this is that the coverage of all Polish verbs using manual sense-level sense annotation is not only not feasible, but also quite possibly is not needed for effective LCM labeling of texts. Hopefully, word usage frequencies follow Zipf distributions and high quality sense-level coverage is important only for frequently used verbs. Many Polish verbs are used only occasionally.

6. Acknowledgements

Work financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education.

7. Bibliographical References

Beukeboom, C., Tanis, M., and Vermeulen, I. (2013). The language of extraversion: Extraverted people talk more abstractly, introverts are more concrete. *Journal of Language and Social Psychology*, 32(2):191–201.

de Montes, L. G., Semin, G. R., and Valencia, J. F. (2003). Communication patterns in interdependent relationships. *Journal of Language and Social Psychology*, 22(3):259–281.

Douglas, K. M. and Sutton, R. M. (2006). When what you say about others says something about you: Language abstraction and inferences about describers' attitudes and goals. *Journal of Experimental Social Psychology*, 42(4):500–508.

Kędzia, P., Piasecki, M., and Orlińska, M. (2015). Word sense disambiguation based on large scale polish clarin heterogeneous lexical resources. *Cognitive Studies / Études cognitives*, 15:269–292.

Maass, A., Salvi, D., Arcuri, L., and Semin, G. R. (1989). Language use in intergroup contexts: the linguistic intergroup bias. *Journal of personality and social psychology*, 57(6):981.

Maass, A., Ceccarelli, R., and Rudin, S. (1996). Linguistic intergroup bias: Evidence for in-group-protective motivation. *Journal of Personality and Social Psychology*, 71(3):512.

Piasecki, M., Szpakowicz, S., and Broda, B. (2009). *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wroclawskiej, Wrocław.

Rogozinska, D. and Wawer, A. (2013). Interpreting or describing? measuring verb abstraction. In *Proceedings of the 2013 IEEE 13th International Conference on Data Mining Workshops*, ICDMW '13, pages 963–966, Washington, DC, USA. IEEE Computer Society.

Rubini, M. and Sigall, H. (2002). Taking the edge off of disagreement: linguistic abstractness and self-presentation to a heterogeneous audience. *European Journal of Social Psychology*, 32(3):343–351.

Schmid, J., Fiedler, K., Semin, G., and Englich, B. (2017). Measuring implicit causality: The linguistic category model. unpublished manuscript.

Semin, G. R. and Fiedler, K. (1988). The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of personality and Social Psychology*, 54(4):558.

Stone, P. J., Dunphy, D. C., Smith, M. S., and Ogilvie, D. M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA.

Wigboldus, D. H., Semin, G. R., and Spears, R. (2000). How do we communicate stereotypes? linguistic bases and inferential consequences. *Journal of personality and social psychology*, 78(1):5.