COGNITIVE STUDIES | ÉTUDES COGNITIVES, 15: 161–175
Warsaw 2015

DOI: 10.11649/cs.2015.013

MAREK MAZIARZ[1,A], **MACIEJ PIASECKI**[1,B], & **STAN SZPAKOWICZ**[2,C]

[1]Department of Computational Intelligence, Wrocław University of Technology, Poland
[2]School of Electrical Engineering & Computer Science, University of Ottawa, Canada
[A]mawroc@gmail.com ; [B]maciej.piasecki@pwr.edu.pl ; [C]szpak@eecs.uottawa.ca

THE SYSTEM OF REGISTER LABELS IN PLWORDNET

Abstract

Stylistic registers influence word usage. Both traditional dictionaries and wordnets assign lexical units to registers, and there is a wide range of solutions. A system of register labels can be flat or hierarchical, with few labels or many, homogeneous or decomposed into sets of elementary features. We review the register label systems in lexicography, and then discuss our model, designed for plWordNet, a large wordnet for Polish. There follows a detailed comparative analysis of several register systems in Polish lexical resources. We also present the practical effect of the adoption of our flat, small and homogeneous system: a relatively high consistency of register assignment in plWordNet, as measured by inter-annotator agreement on a manageable sample. Large-scale conclusions for the whole plWordNet remain to be made once the annotation has been completed, but the experience half-way through this labour-intensive exercise is very encouraging.
Keywords: wordnets; plWordNet; lexical register; large-scale wordnet expansion; inter-annotator agreement

1. Introduction

As many other wordnets, plWordNet is a lexical-semantic network which describes lexical meaning, represented by lexical units,[1] in terms of such lexico-semantic relations as, *e.g.,* hypernymy, hyponymy, meronymy, antonymy, cause and precedence. A wordnet implements the relational paradigm of lexical semantics. LUs are the nodes in a network, *i.e.,* a graph, and the relations define the arcs between pairs of LUs. The network structure is meant to be the principal means for the description of the LUs. Every LU u is characterised by its links (direct) to other LUs which are next linked to further LUs (thus indirectly linked to u), and so on. The wordnet, therefore, describes u by a graph around it, part of the complete network, and this graph imposes restriction on the meaning of u. For example:

[1]The term *lexical unit* will be abbreviated to LU throughout this paper.

- *fura* 4 '≈ (*informal*) a good car' is a hyponym of *samochód osobowy* 1 'a car' which is a hyponym of *samochód* 1 'a motor vehicle';
- *bagażnik* 1 'a luggage compartment' is a meronym of *samochód* 1;
- *gablota* 2 '≈ (*informal*) an expensive large car' is a hyponym of *samochód osobowy* 1;
- and so on.

From the links for '≈ (*informal*) a good car' we can learn that it is a kind of car (which is a kind of vehicle and so on) and it can have parts such as a luggage compartment. We notice, however, that it is a partial description: it does not provide, *e.g.,* a detailed description of situations in which a car can be used, who can drive it and so on. This is an intended effect, because a wordnet is a compromise between the formalisation and the coverage of the description. The wordnet is formalised enough for many applications in Natural Language Engineering, but at the same time its limited formalisation allows for relatively fast work on its construction. As a result, wordnets are among the largest lexical-semantic resources ever built. Their large size and wide coverage are important for their applications.

A hyponym, *e.g.,* '≈ (*informal*) a good car' is more specific than its hypernym 'a car', so the latter can be used in most contexts in which the former is used. The semantic opposition expressed by hyponymy does not explain, however, why the former can be used in all contexts, including formal documents, while the latter is more typical of private conversations or informal texts. This difference can be traced back to the different styles of writing, and cannot be described by the lexico-semantic relations. That is because a style is not a lexical meaning, and cannot be an element of a wordnet, which is a *lexical-semantic* network. We need a different way of introducing limited pragmatic information into the description provided by plWordNet and any wordnet in general.

Our goal is to investigate the use of stylistic registers as a means for expressing selected pragmatic constraints on the lexical meaning described in a wordnet. We want to find the best way of introducing the registers into the wordnet structure, given that they are not relational by nature. We also want to develop a system of stylistic registers for Polish to assist the consistent construction of plWordNet and its future applications.

2. Register label systems in lexicography

Register is usually defined as a language variation stemming from situational characteristic of a communication act. According to Biber (2006):

> "[Register is] any language variety defined by its situational characteristics, including the speaker's purpose, the relationship between speaker and hearer, and the production circumstances".

Halliday (2002, p. 168) defines it thus:

> "Register [is a] functional (diatopic) variation in language."

This language variety includes many aspects of communication, among them formality (*e.g., formal style*), text type (*e.g., literary, poetic*), medium (*e.g., spoken*),

technicality (*e.g., terminology*, *jargon*), frequency (*e.g., rare*), time (*e.g., old-use*, *archaic*), attitude (*e.g., vulgar*, *ironic*), socio-cultural context (*e.g., argot*), normativity (*e.g., non-standard*) or place (*e.g., dialect*, *American English* (Hausmann, 1989)). People can shift between these and many other registers. This style-shifting is triggered by social pressure and requires a higher or lower "amount of attention paid to speech", with spoken colloquial style demanding the least (Milroy & Gordon, 2003). Incapability of such code-switching may be a sign of mental disorders, such as autism (Lyons, 2013).

This extensive and multidimensional variability of language gathered under the umbrella term of *register* (and others, like *style* (Eckert & Rickford, 2001)) may evade precise definition. An example of such problems is the theoretical status of dialect. Dialects are often allocated outside the register list, because of the assumption that one cannot switch from his dialect to the general language or to another dialect in the same way as one jumps into one register from another (Biber & Conrad, 2009, pp. 11–13; Gregory, 1967; Halliday, 2002, pp. 168–169). This common conviction appears to be debunked by research on code-switching, which proves that a dialect could be switched in the same way as style (DeBose, 1992; Trudgill, 1999), and leads to a different register list (Svensén, 2009). Not only do lists of registers vary from one publication to another, but also the boundaries between register types are neither clear nor well established (Bowker, 2013, p. 48). Biber and Conrad (2009, pp. 32–33) claim that the situation is somehow natural, since registers are organised hierarchically and form a continuum; in fact the granularity of register types depends on the researcher's purpose, and on the scope of scientific analysis (Biber & Conrad, 2009). Halliday (2002, p. 169) describes it thus:

> "[Registers] are best thought of as spaces within which the speakers and writers are moving; spaces that may be defined with varying depth of focus (... the register of high school physics textbooks versus the register of natural science), and whose boundaries are in any case permeable, hence constantly changing and evolving."

Multidimensional register systems are arranged into many *scales* with an unmarked/neutral central zone. For example, in Routledge *Dictionary of Lexicography* we note the following scales (Hartmann & James, 2002, after Svensén, 2009):

- the emotiveness scale ("from 'appreciative' through neutral (the unmarked zone) to 'derogatory' and 'offensive' "):
- the formality scale ("from 'elevated' and 'formal' through neutral (the unmarked zone) to 'informal' and 'intimate' ");
- the frequency of occurrence scale ("ranging from 'very frequent' to frequent (the unmarked neutral zone) to 'becoming rare' and 'very rare' ");
- the scale of indigenisation ("from 'foreign' and 'borrowed' through 'assimilated' to native (the unmarked neutral zone)");
- the scale of textuality ("from 'poetic' to 'conversational', with the shared neutral items remaining unmarked");
- the diatopic scale / continuum ("from 'local' or 'provincial' dialects to 'metropolitan' and even 'international' varieties", "[t]he neutral zone of the 'home'

variety (*e.g.,* British English in a British dictionary or American English in an American dictionary) may be left unmarked");
- the diastratic scale ("from neutral (the unmarked zone) to 'demotic' or 'slang' ");
- the dianormative scale ("from 'correct' (the unmarked neutral zone) to 'substandard' or 'illiterate"').

The unmarked/neutral centre of all these scales is the general language (Atkins & Rundell, 2008, p. 498). All other registers are described as marked.

A specific register may be described with regard to some single feature (Biber, 1995, ch. 1.3.1), but real-world registers are fairly complex and must be decomposed in order to find out the underlying simple linguistic features (Maybin & Swann, 2009, pp. 64–65). Biber's model, based on statistical analysis, includes five features (Biber & Conrad, 2009):

(i) 'involved production' ↔ 'informational production',
(ii) 'narrative discourse' ↔ 'non-narrative discourse',
(iii) 'elaborated reference' ↔ 'situation-dependent reference',
(iv) 'overt expression of argumentation',
(v) 'impersonal style' ↔ 'non-impersonal style'.

Buttler and Markowski (1998) proposed an interesting three-dimensional model of lexical registers. Three scales were used: technicality ($\pm t$), formality ($\pm f$), and expressiveness/emotiveness ($\pm e$). Here is the structure of each of the six registers (Buttler & Markowski, 1998, p. 109):

- `common` $[-t, -f, -e]$,
- `literary` $[-t, +f, -e]$,
- `colloquial` $[-t, -f, +e]$,
- `terminological` $[+t, +f, -e]$,
- `professional` $[+t, -f, -e]$,
- `argot` $[+t, -f, +e]$.

Note that it is impossible in this model to combine features $[+f]$ with $[+e]$, so six rather than eight (2^3) possibilities are realised.

Registers are "ways of saying different things" (Halliday, 2002, p. 169), and involve different vocabulary (Biber, 2006). To mark a register of a given word/sense, dictionaries use *register labels* (Svensén, 2009). Register label systems mirror register models, so the difficulties with precise register definitions become a problem for lexicography (Engelking, Markowski, & Weiss, 1989, p. 300).

Indeed, not only is there no consensus what register label system to adopt, but also the very same registers are marked inconsistently (Svensén, 2009, p. 316):

> "Different dictionaries may use different labels, and the categories represented by the labels may have different ranges in different dictionaries. Moreover, there may be differences in labelling practice, so that, in one dictionary, fewer or more lexical items are regarded as formal or informal, correct or incorrect, etc., than in another one (Haussman 1989: 650)."

It is not difficult to find such discrepancies in dictionaries. Let us compare the descriptions of three most frequent senses of the word *clone* in *Cambridge Dictionaries Online* (CDO) (Heacock, 1995–2011) and in *Oxford Dictionaries* (OE) (Simpson, 2013):[2]

1. `General register` 'a plant or animal that has the same genes as the original from which it was produced' (CDO) / `Biology` 'an organism or cell, or group of organisms or cells, produced asexually from one ancestor or stock, to which they are genetically identical' (OE);
2. `Informal` 'someone or something that is very similar to someone or something else' (CDO) / `General register` 'A person or thing regarded as an exact copy of another' (OE);
3. `Computing` 'a computer that operates in a very similar way to the one that it was copied from' (CDO) / `General register` 'a computer designed to simulate exactly the operation of another, typically more expensive, model' (OE).

Clearly, the same state of affairs is present in Polish lexicography (Kurkiewicz, 2007, pp. 29–30; Engelking et al., 1989). In Dubisz (2006), for example, the register system includes over a hundred register labels organised hierarchically, while in Kurkiewicz (2007) the list is shorter.

We prefer to keep the whole system simple. We agree with the editors of the Great Polish Dictionary that "it is better to give less information but base it on reasonably clear criteria" (Kurkiewicz, 2007, p. 30). The next section presents a new system of register labels prepared for plWordNet, very small, well defined, non-hierarchical and with single labels rather than label sequences.

3. A model of register labels in plWordNet

A higher number of stylistic registers allows for more fine-grained distinctions, but it makes assigning LUs to registers more difficult. Inconsistencies between the decisions of different linguists are likely. The similarities among registers are not apparent in a flat structure. A hierarchy of registers could be introduced in order to express generalisations over registers (*e.g.,* specialist registers distinguished but grouped together), but such a solution would only be feasible if there were more registers. The question arises, then, whether a larger number of registers is really needed for plWordNet (or any wordnet, for that matter).

We aim to maintain the high consistency in applying register labels to LUs, so we have decided to build our system only on 11 registers. In order to facilitate the process, the register labels have been arranged into a decision tree presented in Figure 1. A plWordNet editor, in a series of substitution tests, assesses the acceptability of the instances of test expressions. The tree guides her to the final choice of a lexical register label. We will show in section 4 how this ascetic system of registers allows the editors to work with a fair degree of consistency.

[2]Curiously, the dictionaries disagree on the register labels for all three senses, despite the proximity of Cambridge and Oxford...

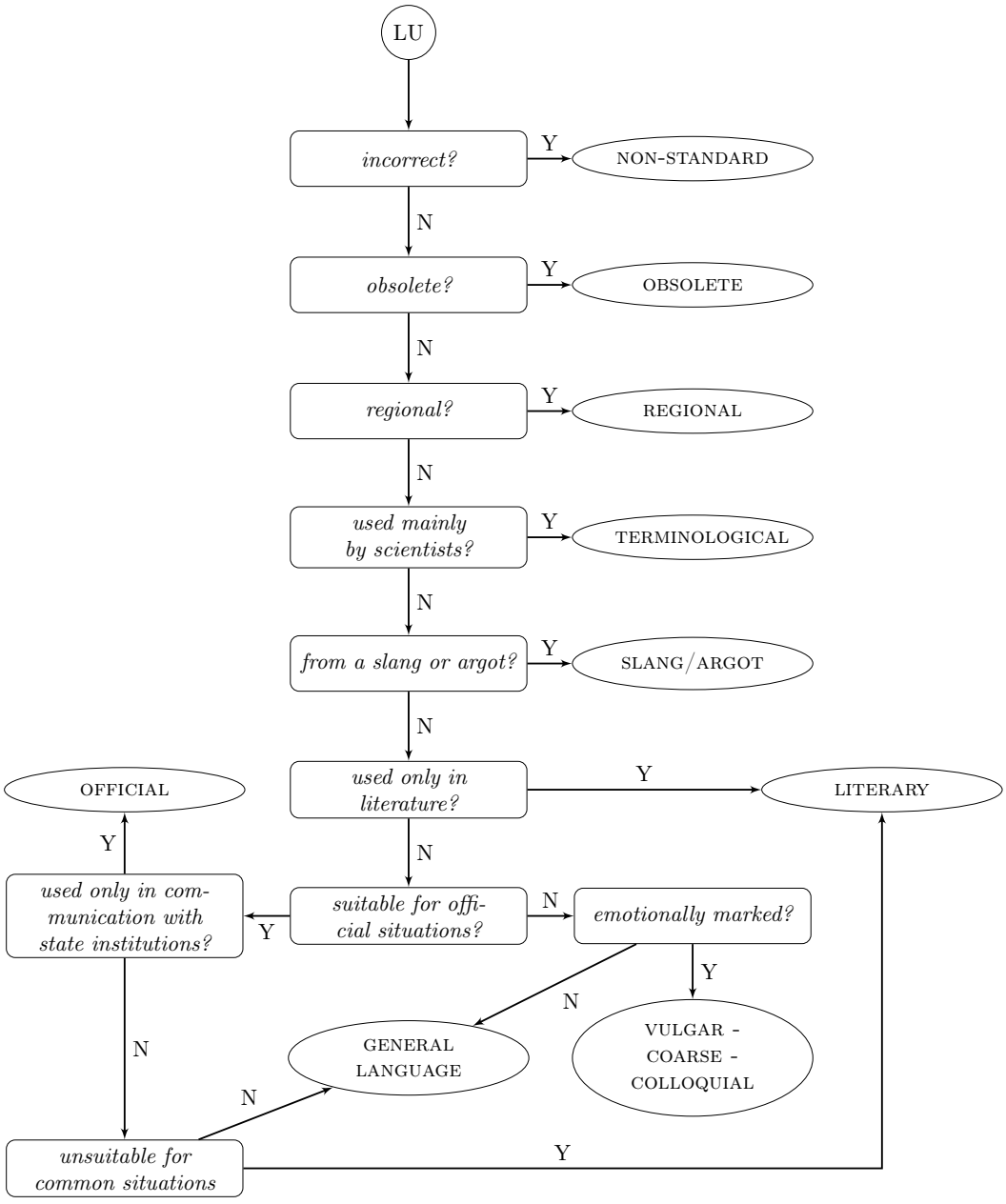


Figure 1: The decision tree for register assignment. The tests for three emotive labels have been conflated.

Registers can be represented — at least for the purpose of analysis — as bundles of primitive semantic features. We consider technicality ($\pm t$), formality ($\pm f$), three levels of expressiveness/emotiveness (according to Engelking et al. (1989): $++\,+e$, $+\,+\,e$, $+e$, $-e$), the status of LU users' community (is it open or closed, like in subcultures, $\pm s$), an exclusively literary character of a LU ($\pm l$), the possibility of using a LU in everyday situation ($\pm u$), and a bureaucratic character of a LU ($\pm b$). The system we have designed includes the following registers:

- NON-STANDARD — we use this register label to mark incorrect but very frequent LUs;
- OBSOLETE — this label marks LUs which are outdated, typically used only by the elderly or (rarely) middle-aged people, as well as in old literature;
- REGIONAL — LUs from a dialect, well known to (but not used by) almost all Poles;
- TERMINOLOGICAL $[+t]$ — LUs used by specialists, scientists, engineers, and generally professionals;
- ARGOT/SLANG $[-t,\ +s]$ — LUs used by a particular closed social group or a small community;
- LITERARY $[-t,\ -s,+l]$, $[-t,\ -s,-l,\ +f,\ -b,\ -u]$ — this label marks high-style vocabulary, especially LUs used only in literature or in speeches;
- OFFICIAL $[-t,\ -s,\ -l,\ +f,\ +b]$ — LUs used on official and formal occasions, mainly in the communication between citizens and representatives of state institutions;[3]
- VULGAR $[-t,\ -s,\ -l,\ -f,\ +\,+\,+e]$ — crude vocabulary, LUs with very restricted acceptable usage;
- COARSE $[-t,\ -s,\ -l,\ -f,\ +\,+\,e]$ — LUs which might be used in a familiar context, but normally not acceptable in other situations;
- COLLOQUIAL $[-t,\ -s,\ -l,\ -f,\ +e]$ — vocabulary used informally, in a free style, but with low acceptability in official situations;
- GENERAL $[-t,\ -s,-l,\ +f,\ -b,\ +u]$, $[-t,\ -s,\ -l,\ -f,\ -e]$ — LUs which could be used virtually in every situation (are common within all styles).

Registers in plWordNet have an important role in shaping the structure of the graph of lexico-semantic relations. In the case of LUs in different registers, we must consider the compatibility of their registers before linking them by a relation, *e.g.,* hyponymy/hypernymy, and thus deciding how they are to be grouped into one synset.

We follow three rules when we link LUs by hyponymy/hypernymy:

1. LU u_1 in the register OBSOLETE, REGIONAL, ARGOT or NON-STANDARD may be a hypernym of LU u_2 if and only if u_2 is in exactly the same register.[4]
2. LU u_1 in the register VULGAR or COARSE may be a hypernym of LU u_2 if and only if u_2 is either in VULGAR or in COARSE.[5]

[3]Such language develops around any bureaucracy.
[4]Each of these registers shows affinity only for itself.
[5]There is affinity between VULGAR and COARSE.

3. The remaining registers may be linked by hyponymy without restrictions.

Synonymy in plWordNet is captured as bidirectional hyponymy (Maziarz, Piasecki, & Szpakowicz, 2013), so very similar rules apply to synomy as well; Table 1 shows the exceptions.

Table 1: Registers allowed in the same synset (+), and those not allowed (−).

	COLL.	GEN.	LIT.	OFF.	TERM.
COLLOQUIAL	+	+	−	−	−
GENERAL	+	+	+	+	+
LITERARY	−	+	+	+	+
OFFICIAL	−	+	+	+	+
TERMINOLOGICAL	−	+	+	+	+

4. Inter-annotator agreement and statistics

At the end of 2013, we constructed the first set of ten register labels. The set was tested and proven useful (Maziarz, Piasecki, Rudnicka, & Szpakowicz, 2014). We then added the 11^{th} register, NON-STANDARD, for the LUs very frequent in Polish but assumed to be incorrect in normative dictionaries. We also conducted a survey. Two of the plWordNet editors applied registers from our set to a random sample of 385 noun LUs taken from plWordNet. The editors were professional linguists, but they had not been trained in register label recognition; they took their guidelines from the decision tree. The distribution of their choices is presented in Table 2; it also shows the statistics of register usage in the newest version of plWordNet (the column 'plWN 2015').

The inter-annotator agreement was determinied by the Cohen's kappa coefficient: the overall agreement was $\kappa = 0.647$ with the confidence interval 0.586-0.722.[6] According to Landis and Koch (1977, p. 165), the confidence interval covers four values of agreement strength: fair – moderate – substantial – almost perfect. We also give kappa values for individual register labels.

A generous rule of thumb in computational linguistics says that only $\kappa \geq 0.8$ guarantees reliable results, and κ in 0.67–0.8 is tolerable.[7] Our result was at the border of the tolerable interval of lower κ (in the terms of confidence intervals). As one can notice, the agreement values between the two annotators were quite good for very frequent registers (TERMINOLOGY: $\kappa = 0.78$, GENERAL REGISTER:

[6]The confidence interval was calculated by a simple percentile bootstrap method (DiCiccio & Efron, 1996; DiCiccio & Romano, 1988) suitable for Cohen's κ (Artstein & Poesio, 2008), $n = 10000$ resamplings, $\alpha = 0.05$.

[7]Reidsma and Carletta (2007) show that this rule of thumb does not always work. Sometimes lower κ makes the results reliable, sometimes even $\kappa \geq 0.8$ does not suffice. That is why in Maziarz et al. (2014) applied to the data a non-parametric test for independence. It proved that neither linguist had a bias. In this paper we also give κ for every category, as suggested in Reidsma and Carletta (2007), so as to inspect the behaviour of agreement across the registers.

Table 2: Inter-rater agreement of two annotators assigning register labels to nouns from plWordNet in 2013, and the frequencies of choices of linguists $F\#1$ and $F\#2$. The label NON-STANDARD was added in 2014. The column 'plWN 2015' contains data from the beginning of 2015.

marking label	Cohen's κ	$F\#1$	%	$F\#2$	%	plWN 2015	%
TERMINOLOGY	0.78	162	42%	146	38%	52 164	59%
GENERAL	0.60	108	28%	113	29%	26 242	29%
LITERARY	0.62	27	15%	33	16%	2 875	3%
COLLOQUIAL	0.52	24	6%	44	11%	3 372	4%
OBSOLETE	0.56	12	3%	9	2%	2 095	2%
COARSE	0.49	9	2%	3	<1%	324	<1%
ARGOT	0.60	5	1%	5	1%	520	<1%
OFFICIAL	−0.01	4	1%	1	<1%	494	<1%
REGIONAL	0.50	3	<1%	1	<1%	832	1%
VULGAR	NA	0	0%	0	0%	65	<1%
NON-STANDARD	NA	0	0%	0	0%	57	<1%
overall	0.647	385	100%	385	100%	89 040	100%

$\kappa = 0.60$), and LITERARY: $\kappa = 0.62$, but lower for less frequent ones (COLLOQUIAL: $\kappa = 0.52$, and OBSOLETE: $\kappa = 0.56$).[8]

The confidence intervals would be narrower if we reduced the number of registers from 11 to 6, having gathered compatible registers into broader bins — see Table 3 and Maziarz et al. (2014). By *compatible* we mean registers with similar definitions (Section 3) and close in the decision tree (Figure 1). After this reduction, the overall $\kappa = 0.72$ with a good confidence interval of $\kappa \in (0.657, 0.785)$. Now all the most frequent registers have sufficiently good kappa values (TERMINOLOGY $\sim$ ARGOT $\sim$ OFFICIAL: $\kappa = 0.77$, GENERAL $\sim$ LITERARY $\sim$ COLLOQUIAL: $\kappa = 0.71$).[9]

With this register labelling system, we began to annotate plWordNet systematically (the column 'plWN 2015' in Table 2). At the time of this writing, 55% of all noun LUs have been assigned registers. We were adding to plWordNet terminological multi-word LUs (mainly from the humanities, social sciences and biology), so the TERMINOLOGY register is overrepresented in the column 'plWN 2015'. Even such an unbalanced but very large sample, however, re-enacts the lead pattern visible in the smaller random sample ('$F\#1$' and '$F\#2$'): TERMINOLOGY is the most frequent register, followed by the GENERAL, LITERARY, COLLOQUIAL and OBSO-

[8] Other registers were too rare to give meaningful values of κ (the confidence intervals were very broad), but we proved statistically that $\kappa > 0$ for all registers except OFFICIAL.

[9] This result shows that disagreements are located in the close neighbourhood in our decision tree (since registers were combined according to their proximity in the tree).

Table 3: Inter-rater agreement of two annotators assigning register labels to nouns from plWordNet in 2013, and the frequencies of choices of linguists $F\#1$ and $F\#2$. The expanded five-label system equates compatible labels, as described in Maziarz et al. (2014). The label NON-STANDARD was added in 2014. The column 'plWN 2015' contains data from the beginning of 2015.

marking label	Cohen's κ	$F\#1$	%	$F\#2$	%	plWN 2015	%
TERMINOLOGY ∼ ARGOT ∼ OFFICIAL	0.77	171	44%	152	40%	53 178	60%
GENERAL ∼ LITERARY ∼ COLLOQUIAL	0.71	190	49%	220	57%	32 489	36%
OBSOLETE	0.56	12	3%	9	2%	2 095	2%
VULGAR ∼ COARSE	0.49	9	2%	3	<1%	324	<1%
REGIONAL	0.50	3	<1%	1	<1%	832	1%
NON-STANDARD	NA	0	0%	0	0%	57	<1%
overall	0.72	385	100%	385	100%	89 040	100%

LETE. Other registers are very rare, summing at most to 2.6% in 'plWN 2015' and '$F\#2$' samples and up to 5.5% in '$F\#1$'.

This high frequency of the terminology register is probably a common feature of large dictionaries. In the third volume of Doroszewski (1958–1962, letters H–K), terminology is the most frequent of all registers (Buttler & Markowski, 1998, pp. 110, 121):[10]

> "First of all, scanty number of lexemes of all three types [i.e., general register – literary – colloquial] is striking as compared to the overall number of dictionary entries. It is settled by the huge amount of terminological and crypto-terminological units in lexical content of the dictionary."

From Buttler & Markowski's analysis of Doroszewski (1958–1962) we know that in vocabulary housed in this dictionary the second rank goes to the obsolete register (2460 occurrences, or 16%, in the 3^{rd} volume). This is so, because Doroszewski's dictionary contains many words from the 19th century and the second half of the 18th century (Piotrowski, 2001, p. 86). (In comparison with this number, it is clear that plWordNet is a *par excellence* contemporary Polish dictionary with its 2% of old-use vocabulary.) Then the most frequent are general register (called *common* by Buttler & Markowski, only 546 occurrences, 371 nominal senses among them), colloquial (216 occurrences, 160 nominal senses) and literary (112, including 58 nominal senses). The proportions of the three lexical layers are shown in Figure 2.

[10]Note that Buttler & Markowski used to apply their own labels to many words from Doroszewski, according to their register model.

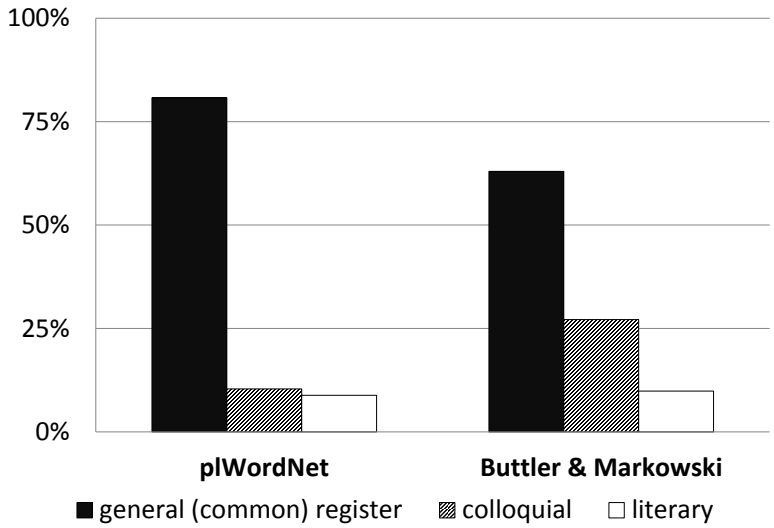


Figure 2: Relative frequencies of three register labels — general, colloquial and literary — in plWordNet, and in Doroszewski (1958–1962) as analysed by Buttler and Markowski (1998).

Both from Buttler & Markowski and from plWordNet we get the same pattern: the most frequent of the three is the general register, followed by the colloquial and the literary.

In Dubisz (2006), the terminology register is also the most common (Table 4, 50%), while the obsolete register is far less frequent (only 3%), as in plWordNet.[11] As we can see from Table 4, the literary, colloquial and general registers are the most frequent ones after terminology.

Putting aside the statistics of terminology and old-use vocabulary, we may focus on three registers which play an important role in the lexical system, *i.e.,* the general (or common) register, the literary register and the colloquial register (Buttler & Markowski, 1998). The distribution of the registers is different in plWordNet and in Buttler & Markowski's model, and that is due to the difference in definitions (Figure 3).

Buttler and Markowski (1998) define the general register with the triple $[-t, -f, -e]$ (Section 2), while in our decision tree (Section 3) the register gets the following feature configurations: $[-t, -s, -l, +f, -b, +u]$, $[-t, -s, -l, -f, -e]$. Because of the semantic feature $+f$ in the former set, the general register of plWordNet has a broader meaning than the common register of Buttler & Markowski. The authors estimate the total population of the common vocabulary at around 5000

[11]The statistics were taken randomly from the dictionary. In the sample of 122 nouns (192 senses) we found 74 unique labels, including 26 complex labels (25 twofold and 1 threefold). Of those 74 labels, 51 represent terminological subregisters, 7 — colloquial, 6 — argot, 4 — literary, 2 — the general register, 2 — the regional register, and 1 each — coarse and official. We have transformed the data into a simpler set, taking into account only the superordinate registers.

Table 4: Register frequencies in a small sample of 122 nouns from Dubisz (2006), 192 senses in total.

Register label	Frequency	%
terminological	95	50%
literary	28	14%
colloquial	26	14%
general	20	10%
argot	8	4%
obsolete	5	3%
official	5	3%
coarse	3	1%
regional	2	1%
sum	192	100%

LUs (ca. 500 LUs × 11 volumes) (Buttler & Markowski, 1998, p. 110). This is much less than in plWordNet: 26 000 in 55% of plWordNet's vocabulary.

The colloquial registers also differ in Doroszewski (1958–1962) and plWordNet. According to Buttler and Markowski (1998) the colloquial register receives the feature set $[-t, -f, +e]$. In plWordNet, the COLLOQUIAL register is simply one of the three registers marked with emotiveness (together with VULGAR and COARSE). Since we single out three levels on the emotiveness scale $[+++e]$, $[++e]$, $[+e]$, in this case the Buttler and Markowski register has a broader meaning than plWordNet's COLLOQUIAL. The literary registers are defined following Buttler and Markowski: $[-t, +f, -e]$, plWordNet: $[-t, -s, +l]$, $[-t, -s, -l, +f, -b, -u]$.

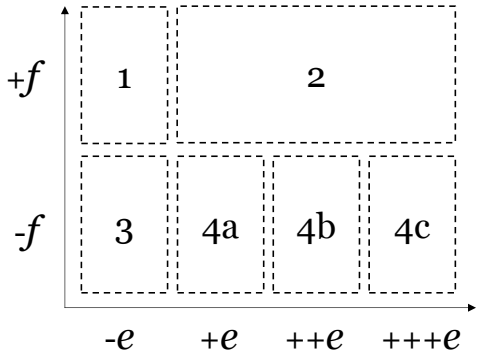
The definitions of the literary registers are also different (Figure 2), mainly because Buttler & Markowski's model disallows features $[+e]$, $[+ + e]$, $[+ + e]$ together with $[+f]$.

5. Concluding remarks

We have proposed an innovative system of stylistic registers for plWordNet, a large Polish wordnet. The system has only 11 registers, is non-hierarchical and always assigns one label to a LU. We have designed a procedure which helps plWordNet editors assign a register label to a given LU. The procedure is summarised in a decision tree accompanied by substitution tests. The editors consult the complete guidelines online.[12]

The register labels significantly affect the structure of plWordNet, because hyponymy/hypernymy and synonymy only link LUs whose registers show affinity for each other.

[12] `http://tinyurl.com/plWN-registers`



Buttler & Markowski (1998)

1 – literary	2 – *null*
3 – general (common)	4a, 4b, 4c – colloquial

plWordNet

1 – general & literary	2 – general & literary
3 – general	4a – colloquial
	4b – coarse
	4c – vulgar

Figure 3: Differences in the definitions of the general register and the colloquial and literary registers between Buttler and Markowski (1998) and plWordNet with regard to the register scales of formality $\{-f, +f\}$ and emotiveness $\{-e, +e, ++e, +++e\}$. The plWordNet general register has a broader extension than the common register in Buttler & Markowski's model, while their colloquial register is a superordinate term for colloquial — coarse — vulgar in plWordNet. Field 2 is a forbidden area in their model: that is why the literary registers have different definitions. All definitions from plWordNet were "translated" into the semantic description language of Buttler & Markowski; we had to project our multidimensional definitions onto a two-dimensional description in terms of formality and emotiveness.

We have examined the consistency of the procedure and found it reasonable. We measure it as inter-annotator agreement, obtaining sufficiently high values of Cohen's kappa. Bundling three groups of compatible labels gives a system with only six categories, and the kappa values for that system are even higher.

Finally, we have compared the statistics: plWordNet half-way through a complete annotation; the Universal Dictionary of Polish; and Buttler & Markowski's model. The distribution of labels is fairly similar, but details differ due to the differences in the underlying register systems.

References

Artstein, R. & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596. http://dx.doi.org/10.1162/coli.07-034-R2

Atkins, B. T. S. & Rundell, M. (2008). *The Oxford guide to practical lexicography.* Oxford: Oxford University Press.

Biber, D. & Conrad, S. (2009). *Register, genre, and style.* Cambridge: Cambridge University Press. Retrieved from http://dx.doi.org/10.1017/CBO9780511814358

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison.* Cambridge: Cambridge University Press.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers.* Amsterdam: John Benjamins. Retrieved from http://dx.doi.org/10.1075/scl.23

Bowker, J. (2013). Variation across spoken and written registers in internal corporate communication: Multimodality and blending in evolving genres. In J. Bamford, S. Cavalieri, & G. Diani (Eds.), *Variation and change in spoken and written discourse* (pp. 47–64). Amsterdam: John Benjamins. Retrieved from http://dx.doi.org/10.1075/ds.21.08bow

Buttler, D. & Markowski, A. (1998). Słownictwo wspólnoodmianowe, książkowe i potoczne współczesnej polszczyzny. *Język a kultura*, (1), 179–203.

DeBose, C. E. (1992). Codeswitching: Black English and standard English in the African-American linguistic repertoire. *Journal of Multilingual and Multicultural Development*, *13*(1–2), 157–167. http://doi.org/10.1080/01434632.1992.9994489

DiCiccio, T. J. & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, *11*(3), 189–212. http://dx.doi.org/10.1214/ss/1032280214

DiCiccio, T. J. & Romano, J. P. (1988). A review of bootstrap confidence intervals. *Journal of the Royal Statistical Society. Series B (Methodological)*, *50*(3), 338–354.

Doroszewski, W. (Ed.). (1958–1962). *Słownik języka polskiego* (Vol. 3). Warszawa: PWN.

Dubisz, S. (2006). Wstęp. In S. Dubisz (Ed.), *Uniwersalny słownik języka polskiego PWN. Wersja 3.0* [CD]. Warszawa: Wydawnictwo Naukowe PWN.

Eckert, P. & Rickford, J. (2001). *Style and sociolinguistic variation.* Cambridge: Cambridge University Press.

Engelking, A., Markowski, A., & Weiss, E. (1989). Kwalifikatory w słownikach — próba systematyzacji. *Poradnik Językowy*, (5), 300–309.

Gregory, M. (1967). Aspects of varieties differentiation. *Journal of Linguistics*, *3*(02), 177–197. http://dx.doi.org/10.1017/S0022226700016601

Halliday, M. A. K. (2002). The construction of knowledge and value in the grammar of scientific discourse: With reference to Charles Darwin's The origin of species (1990). In J. Webster (Ed.), *Collected works of M.A.K. Halliday* (Vol. 2: *Linguistic studies of text and discourse*, pp. 168–193). London: Continuum.

Hartmann, R. R. K. & James, G. (2002). *Dictionary of lexicography.* London: Routledge.

Hausmann, F. J. (1989). Die Markierung im allgemeinen einsprachigen Wörterbuch: Eine Übersicht. In F. J. Hausmann, O. Reichmann, H. E. Wiegand, & L. Zgusta (Eds.), *Wörterbücher: Ein internationales Handbuch zur Lexikographie* (Vol. 5.1, pp. 649–657). New York: De Gruyter.

Heacock, P. (Ed.). (1995–2011). *Cambridge dictionaries online.* Cambridge: Cambridge University Press.

Kurkiewicz, J. (2007). Kwalifikatory w Wielkim słowniku języka polskiego. In P. Żmigrodzki & R. Przybylska (Eds.), *Nowe studia leksykograficzne*. Kraków: Wydawnictwo Lexis.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. http://dx.doi.org/10.2307/2529310

Lyons, M. (2013). Register variation. In F. R. Volkmar (Ed.), *Encyclopedia of autism spectrum disorders* (p. 2534). New York: Springer. http://www.springerlink.com/index/10.1007/978-1-4419-1698-3_983

Maybin, J. & Swann, J. (2009). *The Routledge Companion to English Language Studies*. New York: Routledge.

Maziarz, M., Piasecki, M., & Szpakowicz, S. (2013). The chicken-and-egg problem in wordnet design: Synonymy, synsets and constitutive relations. *Language Resources and Evaluation*, *47*(3), 769–796. http://dx.doi.org/10.1007/s10579-012-9209-9

Maziarz, M., Piasecki, M., Rudnicka, E., & Szpakowicz, S. (2014). Registers in the system of semantic relations in plWordNet. In *Proceedings of 7th International Global Wordnet Conference* (pp. 330–337).

Milroy, L. & Gordon, J. (2003). *Sociolinguistics: Method and interpretation*. Cambridge, MA: Blackwell Publishing Ltd. Retrieved from http://dx.doi.org/10.1002/9780470758359

Piotrowski, T. (2001). *Zrozumieć leksykografię*. Warszawa: PWN.

Reidsma, D. & Carletta, J. (2007). Reliability measurement without limits. *Computational Linguistics*, *1*(1), 1–8.

Simpson, J. (2013). *Oxford English Dictionary*. Oxford: Oxford University Press. Retrieved from public.oed.com/

Svensén, B. (2009). *A handbook of lexicography: The theory and practice of dictionary-making*. New York: Cambridge University Press.

Trudgill, P. (1999). Standard English: What it isn't. In T. Bex & R. J. Watts (Eds.), *Standard English: The widening debate* (pp. 117–128). London: Routledge.

Acknowledgment

This work was supported by a grant from the Polish Ministry of Science and Higher Education, a program in support of scientific units involved in the development of a European research infrastructure for the humanities and social sciences in the scope of the consortia CLARIN ERIC and ESS-ERIC, 2015–2016.

The authors declare that they have no competing interests.

The authors' contribution was as follows: concept of the study: Marek Maziarz, Maciej Piasecki, Stan Szpakowicz; data analyses: Marek Maziarz, Maciej Piasecki, Stan Szpakowicz; the writing: Marek Maziarz, Maciej Piasecki, Stan Szpakowicz.

Publisher: Institute of Slavic Studies, PAS, University of Silesia & The Slavic Foundation