Corpus-based Wordnet Development and plWordNet as a Relational Semantic Dictionary

Maciej Piasecki, Agnieszka Dziob
Wrocław University of Science and Technology
G4.19 Research Group
maciej.piasecki@pwr.wroc.pl

2017-02-14

Politechnika Wrocławska

CL G 4.19

CLARIN-PL
Common Language Resources and Technology Infrastructure

Plan

- Wordnet as a dictionary and a useful language resource
- Wordnet is not enough - a system of resources
- Corpus-based wordnet development process
- plWordNet model
  - Synset definition
  - constitutive relations and features
- Semi-automated wordnet expansion – tools for lexicographers
- plWordNet relations – procedural definitions
- Non-relational elements of the wordnet structure
- plWordNet in use
- Conclusions

Background

- Is a wordnet a useful language resource?
- Not many wordnets have influence comparable to Princeton WordNet
  - … but almost none of them come close to WordNet’s size and coverage
  - most of them have been translated, one way or another, from WordNet
- So, the answer is Yes
  - if only a wordnet is large enough
  - has good coverage,
  - and is close to language data coming from corpora
- A story 12 years of **plWordNet** (Polish name: ***Słowosieć***)

plWordNet (Słowosieć): Goal

To build a wordnet which provides a faithful and comprehensive description of the system of Polish lexical semantics

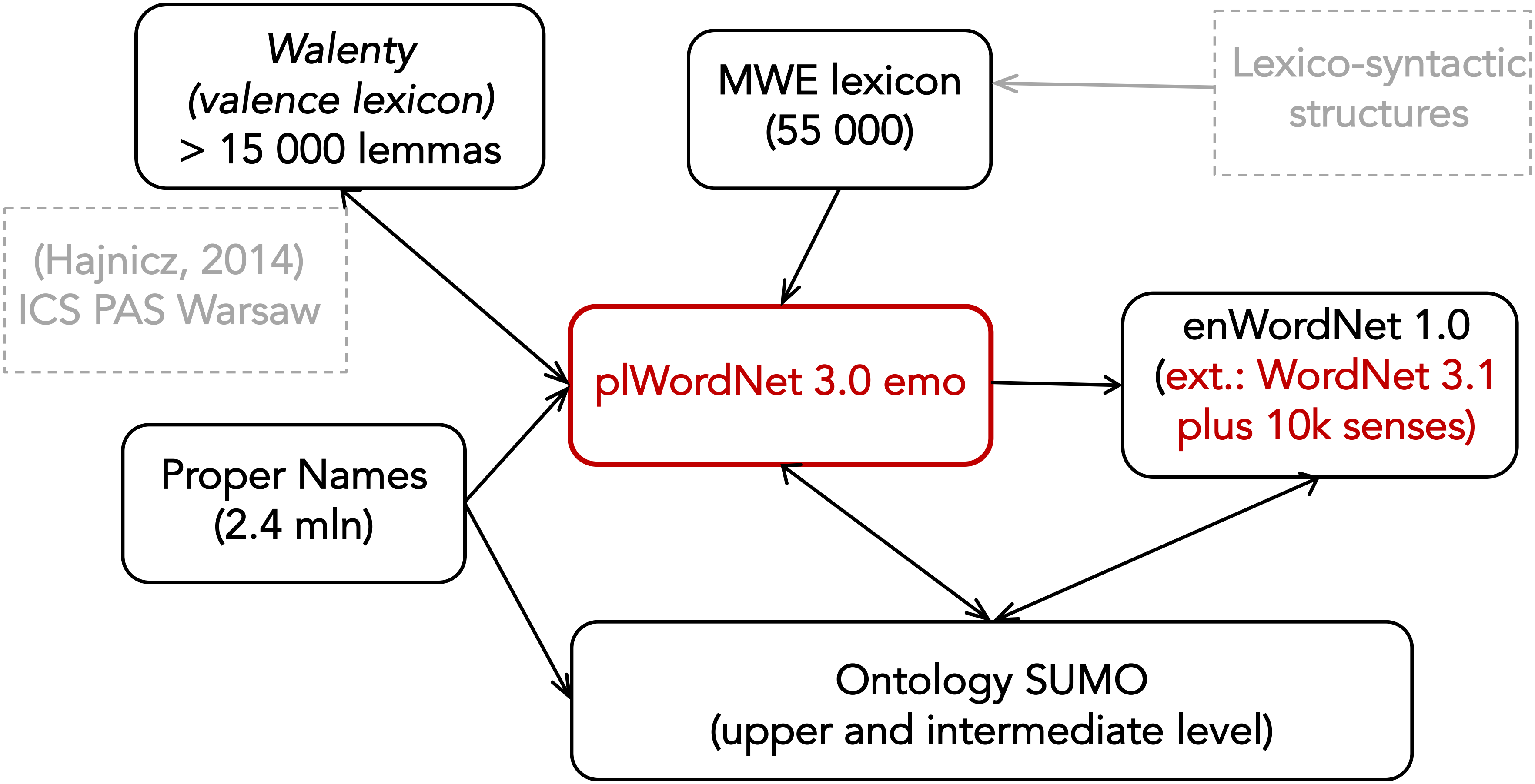
- its structure should represent accurately the lexico-semantic relations between lexical meanings in Polish
- and be motivated only by observations derived from Polish language data
- any form of translation from wordnets for other languages was excluded
- a resource with good coverage with respect to lemmas, word senses and instances of lexico-semantic relations
- in close correspondence to language data collected from very large corpora

A Wordnet is not Enough

- Morphological dictionary (Woliński, 2014)
- Lexicon of Multiword Expressions (structurally described) (Kurc et al., 2012)
- Lexico-semantic resources – lexical meanings
  - plWordNet 4.0 emo (Słowosieć)
  - Syntactic-semantic valency lexicon – Walenty (Przepiórkowski et al., 2014) IPI PAN
  - enWordNet 1.0 – a significant expansion of WordNet 3.1
  - Mapping of plWordNet onto enWordNet
- Knowledge resources
  - NELexicon 2.0 – a large lexicon of Proper Names
  - Mapping of plWordNet onto SUMO Ontology
  - Mapping of plWordNet onto Wikipedia articles (partial)

Result

- A complex system of lexico-semantic resources (Maziarz et al. 2016)



Walenty
(valence lexicon)
> 15 000 lemmas

MWE lexicon
(55 000)

Lexico-syntactic structures

(Hajnicz, 2014)
ICS PAS Warsaw

plWordNet 3.0 emo

enWordNet 1.0
(ext.: WordNet 3.1 plus 10k senses)

Proper Names
(2.4 mln)

Ontology SUMO
(upper and intermediate level)

Corpus-based wordnet development process

- A large text corpus is primary data
  - Lemmas (starting with the most frequent)
  - Examples of use and senses
- Language tools and systems support corpus exploration
  - simple, e.g. concordances
  - advanced – extraction of semantic similarity, relations, sense clusters
  - combined – semi-automated wordnet expansion (Paintball algorithm, RANLP 2013)
- Process
  - systematic extraction of lemmas, acquisition of lexico-semantic knowledge, generation of suggestions, decisions of editors
  - supported by: dictionaries, encyclopaedias, intuition, team

plWordNet model

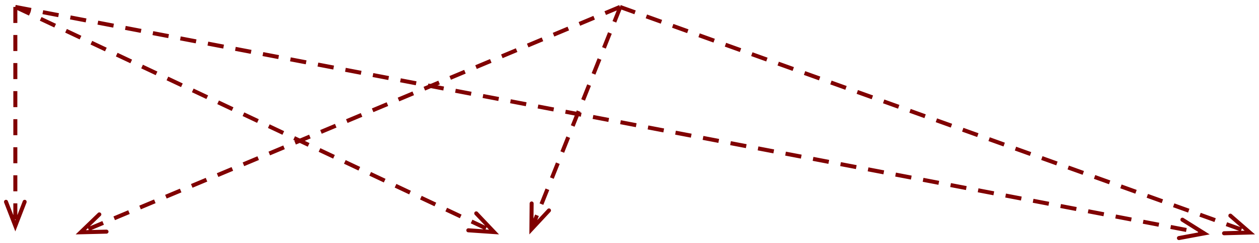
- Corpora contain words, senses discernible by context, not sets of synonyms
- Lexical unit (LU)
  - a triple: <part of speech, lemma, sense id>
  - the basic building block in plWordNet, belongs to one synset
- Synset - a group of lexical units which share
  - lexico-semantic *constitutive relations,* e.g. hyper/hyponymy, mero/holonymy
  - and *constitutive features*: stylistic register, aspect, and semantic classes for adjectives and verbs
- A relation between two synsets is a shorthand for sharing relations between lexical units
- ***Minimal Commitment Principle***

Synset

- Example

  {*wzór 1 `paragon’ ,wzorzec 2* `pattern ‘,…}

  —hypernym→



  {idol 1 `idol ’, bożyszcze 1 `~idol ’ , gwiazdor 1 ~ `star ’}

- Synset as a notational convention
  - for a group of lexical units sharing certain constitutive relations
- What are wordnet constitutive relations?
- Are relations enough to define synsets?

Constitutive relations

- Required properties
  - well-established in linguistics
    - good understanding (e.g. paradigmatic relations)
    - existing descriptions
  - definable with sufficient specificity
  - and useful in generalisation
    - relatively frequent
    - should describe sets of lexical units systematically – a sharing factor
- Level of generalisation of a wordnet vs selection of the constitutive relations

Constitutive features

- Wordnet structure as a basis for acceptable conclusions
  - lack of formal definitions
  - some conclusions based on properties of relations, e.g. transitivity of hypernymy
- Additional constraints on the relation definitions
  - meta-conditions
  - obligatory and built into the relation definitions
- plWordNet: stylistic registers, semantic verb classes and aspect

Corpus-based Wordnet Development

- Limited resources at the starting point
    - translation ruled out & no electronic monolingual dictionaries to leverage
- Schema
    1. A large corpus built from available sources
    2. Extraction of lemma frequency list
    3. Selection of new lemmas
    4. Building a Measure of Semantic Relatedness
    5. MSR-based clustering new lemma into packages
    6. Extraction of knowledge sources
    7. Wordnet editing supported by tools
        - Semi-automated wordnet expansions
        - Semantic exploration of corpora
        - Consulting traditional linguistic resources
    8. Linguistic work management

Corpus-based Wordnet Development

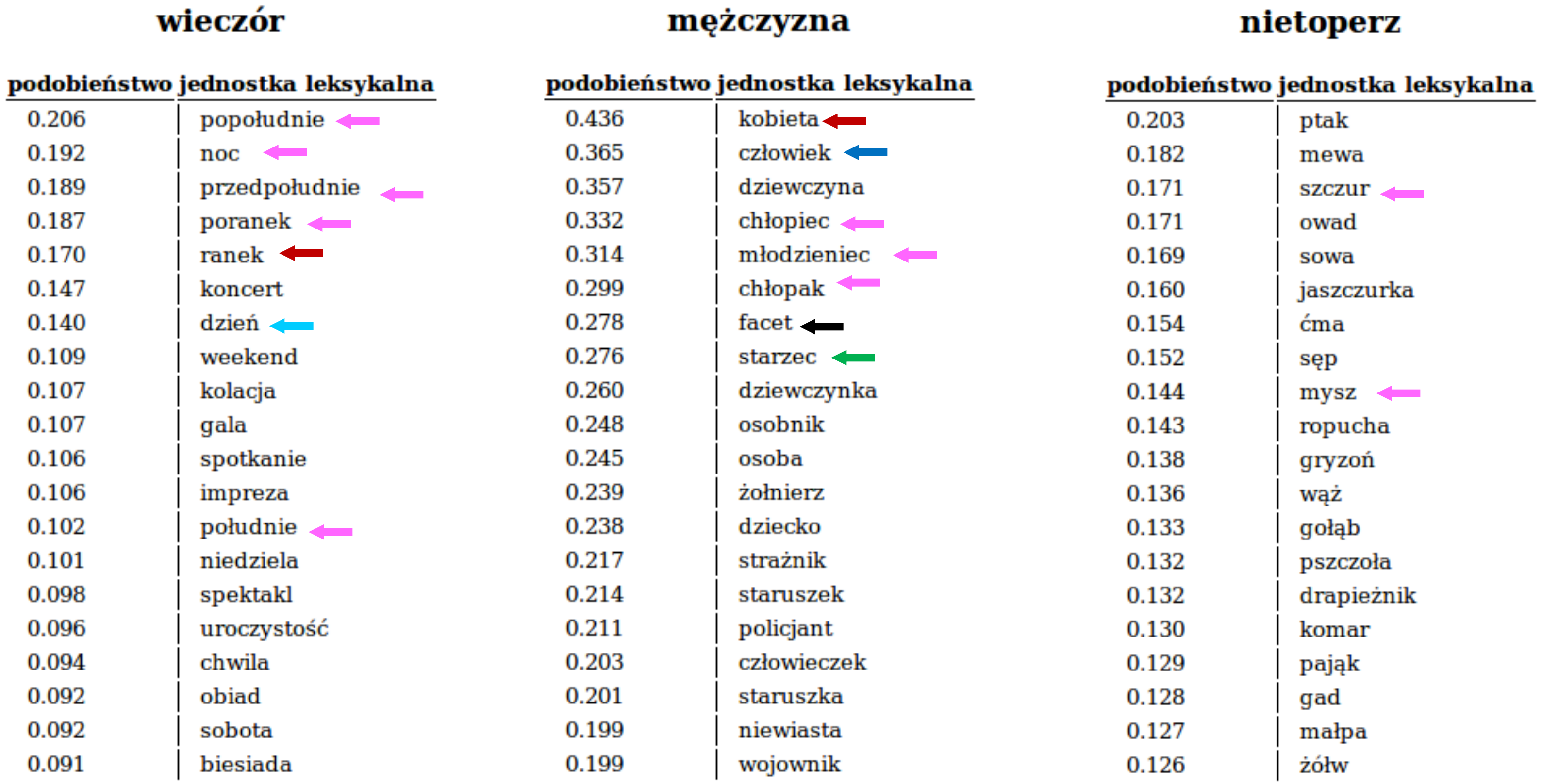
1. plWordNet Merged Corpus
    - available Polish corpora:
        - Corpus IPI PAN
        - Rzeczypospolita Corpus
        - Wikipedia (2015)
        - Texts on open licence
    - Text collected from Internet
        - larger texts
        - Max. 20% tokens not recognised by Morfeusz analyser
    - The version 7.0: ~ 2 billion tokens
    - The version 10.0: >4 billion tokens (for plWordNet 4.0)

2.&3. Extraction of lemma frequency list and Selection

- from the morpho-syntactically tagged and lemmatised corpus
- necessary manual filtering
- 7 000-9 000 new lemmas of a PoS in focus per iteration

plWordNet development proces

4. Measure of Semantic Relatedness generation: SuperMatrix or word2vec
5. MSR-based clusters of lemmas (up to 200) -> assignment of task for linguists



wieczór

podobieństwo	jednostka leksykalna
0.206	popołudnie 🟪
0.192	noc 🟪
0.189	przedpołudnie 🟪
0.187	poranek 🟪
0.170	ranek 🟥
0.147	koncert
0.140	dzień 🩵
0.109	weekend
0.107	kolacja
0.107	gala
0.106	spotkanie
0.106	impreza
0.102	południe 🟪
0.101	niedziela
0.098	spektakl
0.096	uroczystość
0.094	chwila
0.092	obiad
0.092	sobota
0.091	biesiada

mężczyzna

podobieństwo	jednostka leksykalna
0.436	kobieta 🟥
0.365	człowiek 🟦
0.357	dziewczyna
0.332	chłopiec 🟪
0.314	młodzieniec 🟪
0.299	chłopak 🟪
0.278	facet ⬛
0.276	starzec 🟩
0.260	dziewczynka
0.248	osobnik
0.245	osoba
0.239	żołnierz
0.238	dziecko
0.217	strażnik
0.214	staruszek
0.211	policjant
0.203	człowieczek
0.201	staruszka
0.199	niewiasta
0.199	wojownik

nietoperz

podobieństwo	jednostka leksykalna
0.203	ptak
0.182	mewa
0.171	szczur 🟪
0.171	owad
0.169	sowa
0.160	jaszczurka
0.154	ćma
0.152	sęp
0.144	mysz 🟪
0.143	ropucha
0.138	gryzoń
0.136	wąż
0.133	gołąb
0.132	pszczoła
0.132	drapieżnik
0.130	komar
0.129	pająk
0.128	gad
0.127	małpa
0.126	żółw

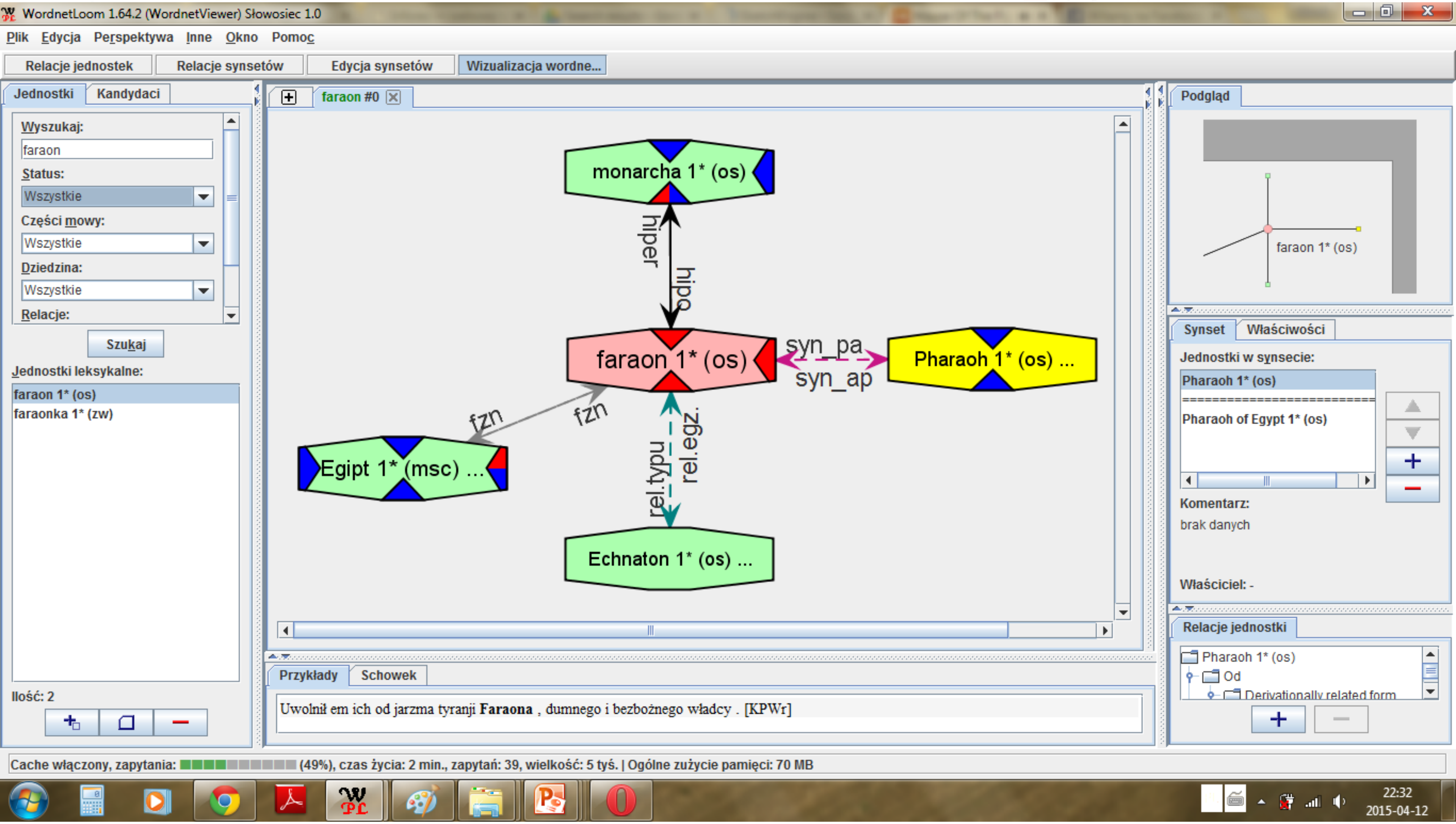
🟥 antonym 🟦 hypernym 🟩 hyponym 🟪 co-hyponym

⬛ closely related 🩵 holonym

Corpus-based Wordnet Development

6. Extraction of knowledge sources
    - Measure of Semantic Relatedness
    - relation instances (hypernymy) extracted by manually constructed lexico-syntactic patterns
    - relation instances extracted by more generic patterns developed in a remotely controlled process
    - ML-based classifier for relation instances
7. Semi-automated wordnet expansions
    - Generation of suggestions for the placement of new lemmas in the wordnet structure
    - Presented for final editing decisions by linguists
    - WordnetWeaver - an extension to WordnetLoom

WordnetLoom – Wordnet Editor



Wyszukaj: faraon

Status: Wszystkie

Części mowy: Wszystkie

Dziedzina: Wszystkie

Relacje:

Jednostki leksykalne:
faraon 1* (os)
faraonka 1* (zw)

Ilość: 2

Jednostki w synsecie:
Pharaoh 1* (os)
==========================
Pharaoh of Egypt 1* (os)

Komentarz:
brak danych

Właściciel: -

Pharaoh 1* (os)
Od
Derivationally related form

Uwolnił em ich od jarzma tyranji **Faraona** , dumnego i bezbożnego władcy . [KPWr]

Cache włączony, zapytania: (49%), czas życia: 2 min., zapytań: 39, wielkość: 5 tyś. | Ogólne zużycie pamięci: 70 MB

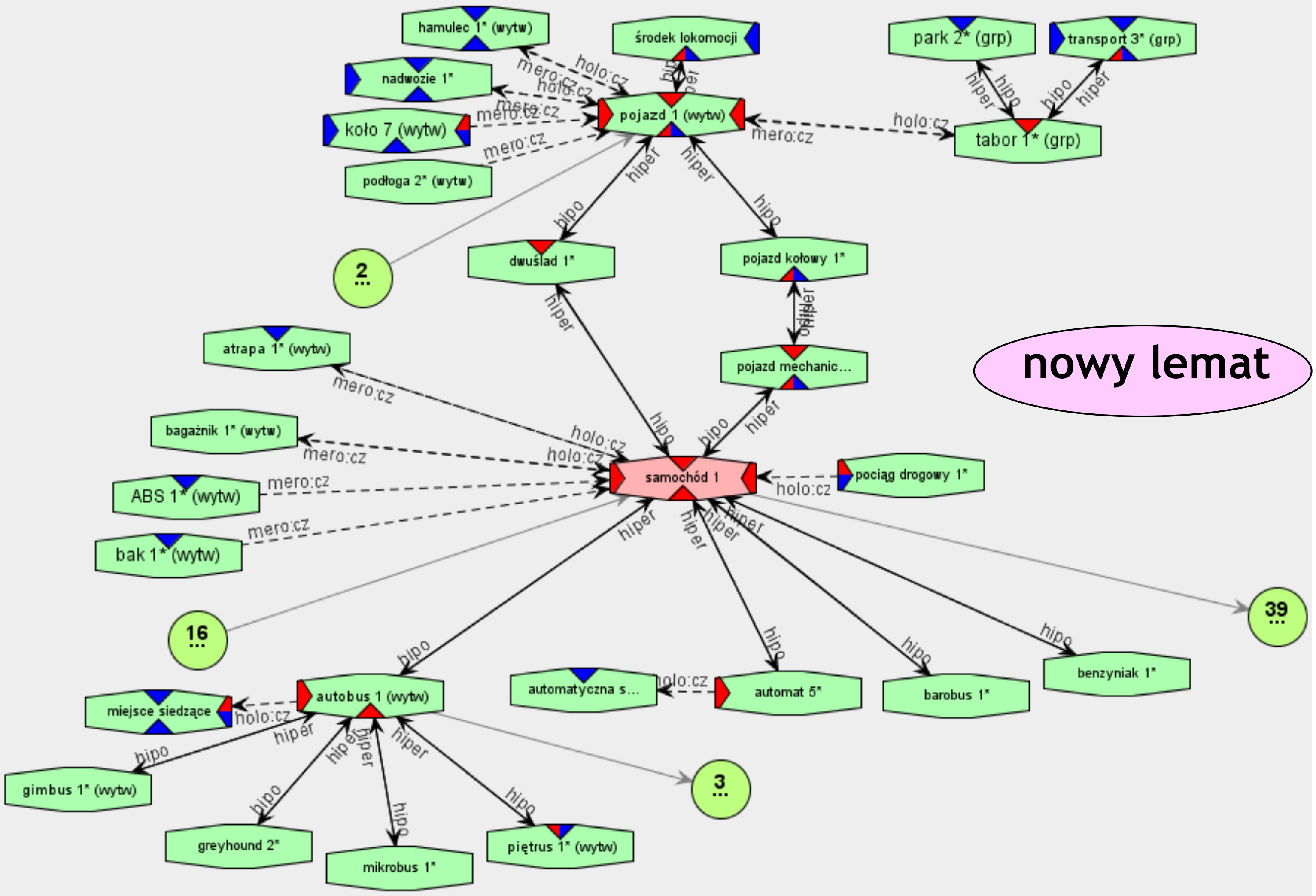
22:32
2015-04-12

Cf (Piasecki et al., 2013b)

Paintball: Knowledge sources

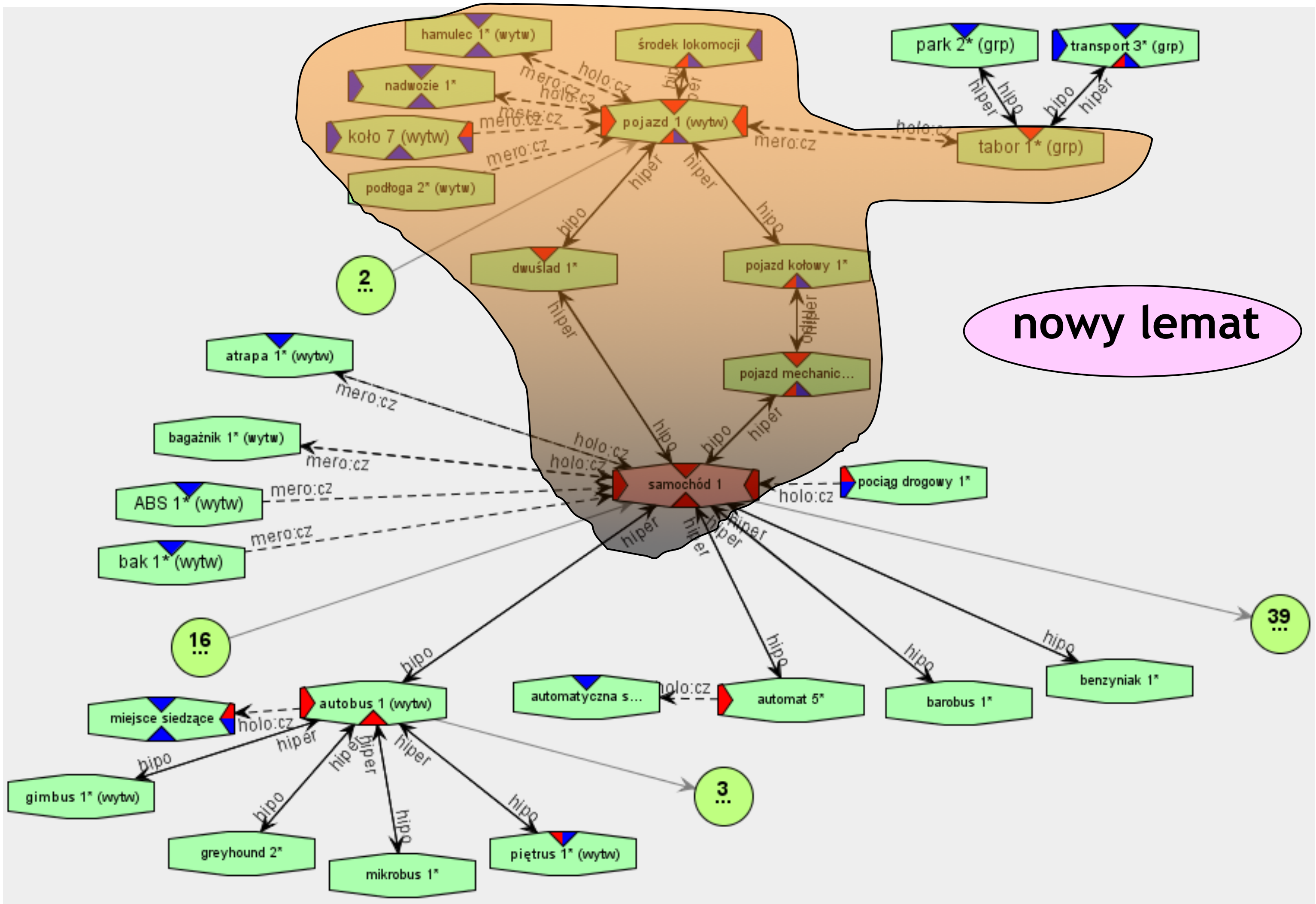
- Methods
  - Measure of Semantic Relatedness
  - Lexico-syntactic Patterns
    - specific – manually constructed
    - generic – automatically extracted
  - Classifiers based on Machine Learning
- Only some of them produce probability values
- Results: heterogeneous, partial, and imperfect – substantial error level

Paintball Metaphor: initial state



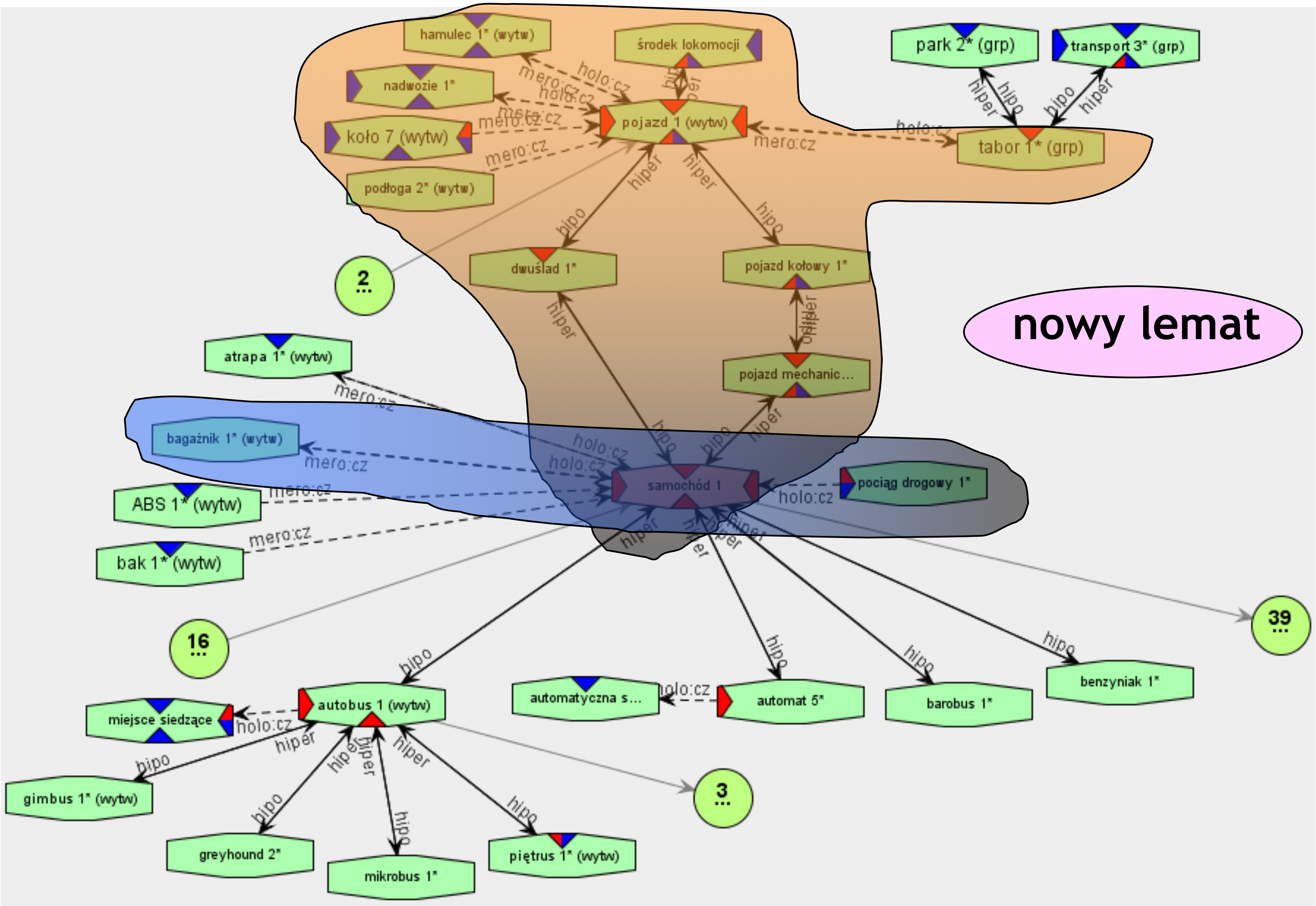
nowy lemat

Paintball Metaphor: hits from the knowledge sources

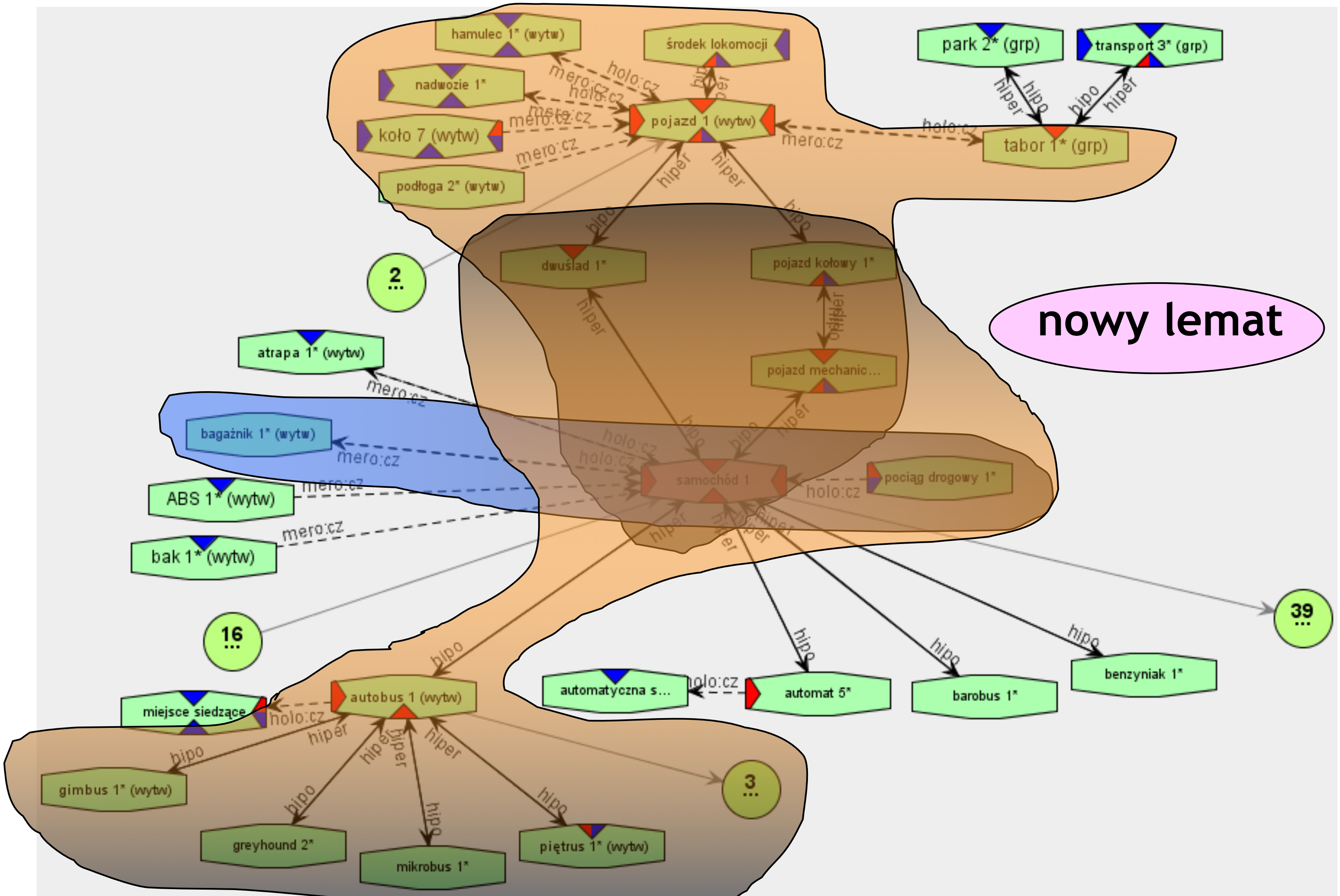


nowy lemat

Paintball Metaphor: hits from the knowledge sources

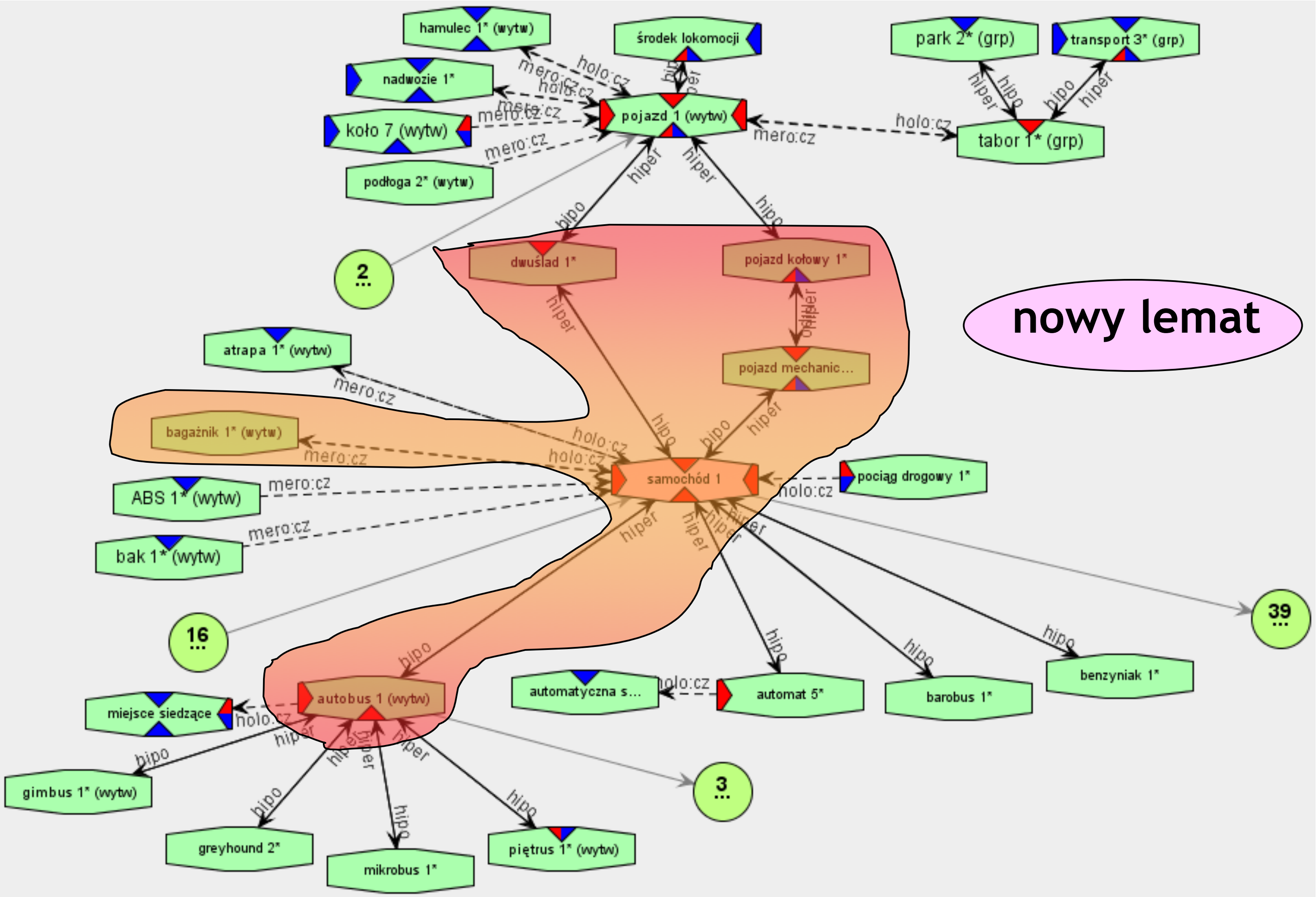


Paintball Metaphor: hits from the knowledge sources



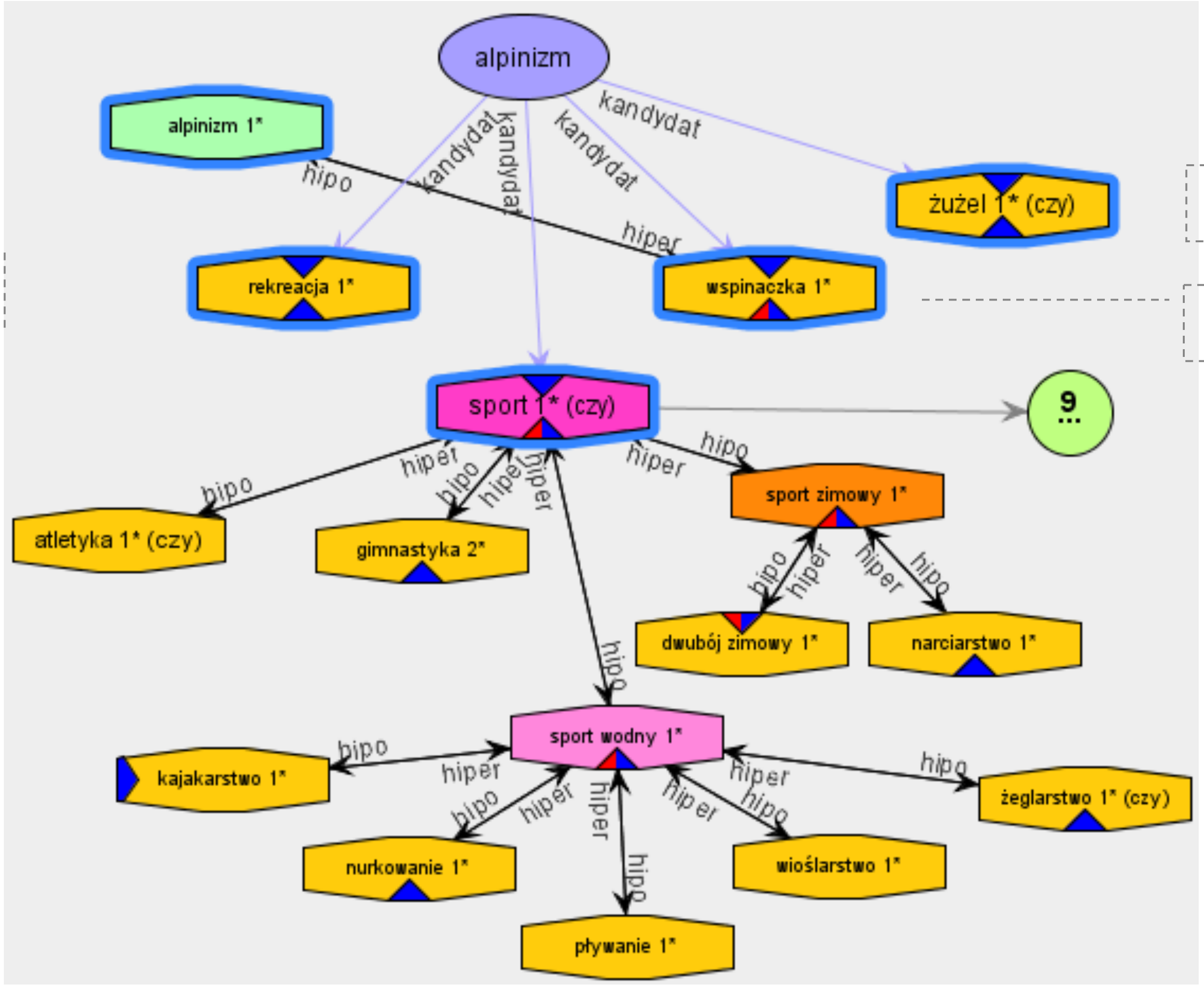
nowy lemat

Paintball Metaphor: attachment area



nowy lemat

WordnetWeaver - Semi-automated Wordnet Expansion



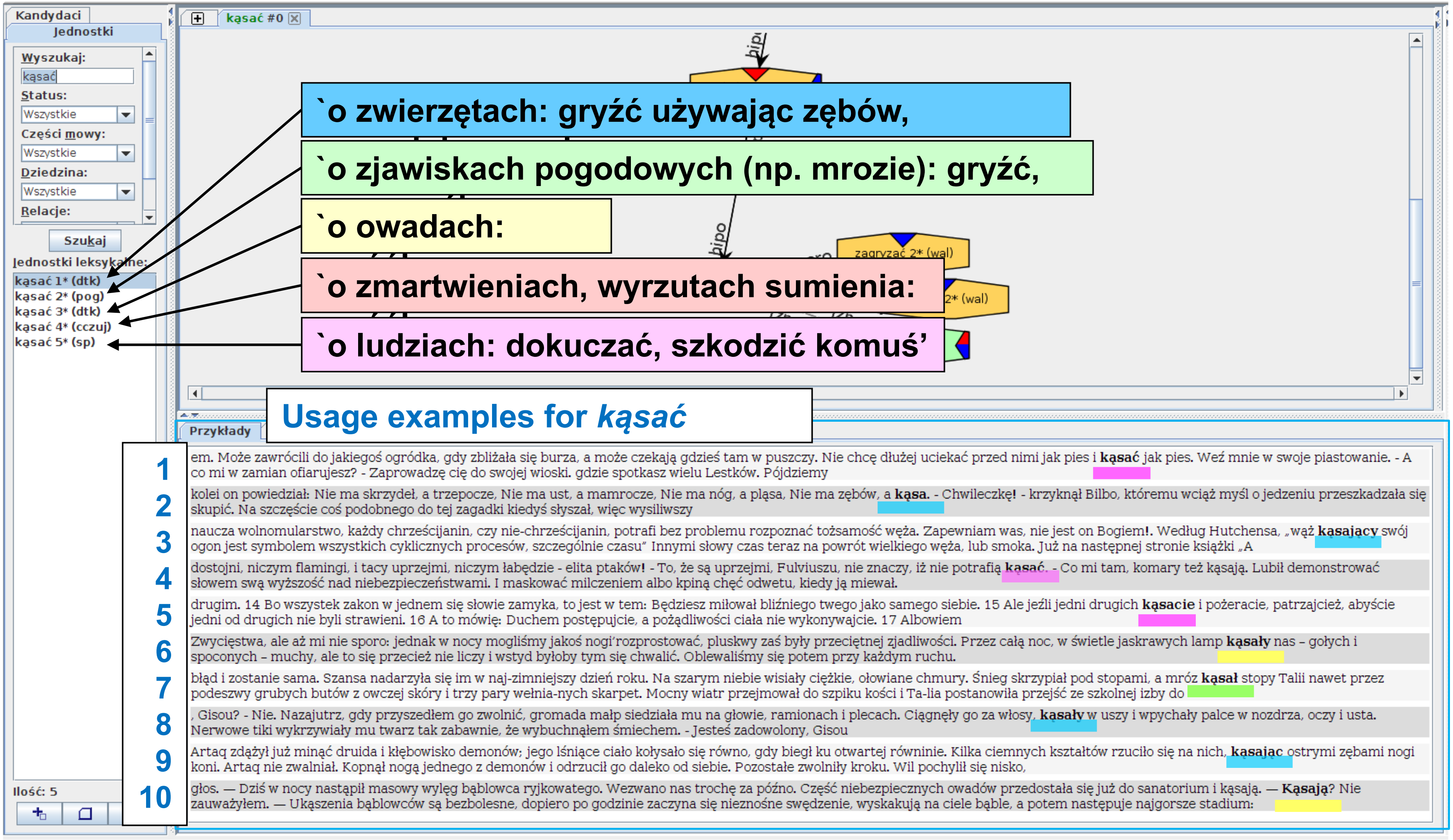
recreation

speedway

climbing

Suggestions generated by **Paintball algorithm**

Semantic exploration of corpora: usage examples (LexCSD)



`o zwierzętach: gryźć używając zębów,

`o zjawiskach pogodowych (np. mrozie): gryźć,

`o owadach:

`o zmartwieniach, wyrzutach sumienia:

`o ludziach: dokuczać, szkodzić komuś'

Usage examples for *kąsać*

1 em. Może zawrócili do jakiegoś ogródka, gdy zbliżała się burza, a może czekają gdzieś tam w puszczy. Nie chcę dłużej uciekać przed nimi jak pies i **kąsać** jak pies. Weź mnie w swoje piastowanie. - A co mi w zamian ofiarujesz? - Zaprowadzę cię do swojej wioski. gdzie spotkasz wielu Lestków. Pójdziemy

2 kolei on powiedział: Nie ma skrzydeł, a trzepocze, Nie ma ust, a mamrocze, Nie ma nóg, a pląsa, Nie ma zębów, a **kąsa.** - Chwileczkę! - krzyknął Bilbo, któremu wciąż myśl o jedzeniu przeszkadzała się skupić. Na szczęście coś podobnego do tej zagadki kiedyś słyszał, więc wysiliwszy

3 naucza wolnomularstwo, każdy chrześcijanin, czy nie-chrześcijanin, potrafi bez problemu rozpoznać tożsamość węża. Zapewniam was, nie jest on Bogiem!. Według Hutchensa, „wąż **kąsający** swój ogon jest symbolem wszystkich cyklicznych procesów, szczególnie czasu" Innymi słowy czas teraz na powrót wielkiego węża, lub smoka. Już na następnej stronie książki „A

4 dostojni, niczym flamingi, i tacy uprzejmi, niczym łabędzie - elita ptaków! - To, że są uprzejmi, Fulviuszu, nie znaczy, iż nie potrafią **kąsać.** - Co mi tam, komary też kąsają. Lubił demonstrować słowem swą wyższość nad niebezpieczeństwami. I maskować milczeniem albo kpiną chęć odwetu, kiedy ją miewał.

5 drugim. 14 Bo wszystek zakon w jednem się słowie zamyka, to jest w tem: Będziesz miłował bliźniego twego jako samego siebie. 15 Ale jeźli jedni drugich **kąsacie** i pożeracie, patrzajcież, abyście jedni od drugich nie byli strawieni. 16 A to mówię: Duchem postępujcie, a pożądliwości ciała nie wykonywajcie. 17 Albowiem

6 Zwycięstwa, ale aż mi nie sporo: jednak w nocy mogliśmy jakoś nogi'rozprostować, pluskwy zaś były przeciętnej zjadliwości. Przez całą noc, w świetle jaskrawych lamp **kąsały** nas – gołych i spoconych – muchy, ale to się przecież nie liczy i wstyd byłoby tym się chwalić. Oblewaliśmy się potem przy każdym ruchu.

7 błąd i zostanie sama. Szansa nadarzyła się im w naj-zimniejszy dzień roku. Na szarym niebie wisiały ciężkie, ołowiane chmury. Śnieg skrzypiał pod stopami, a mróz **kąsał** stopy Talii nawet przez podeszwy grubych butów z owczej skóry i trzy pary wełnia-nych skarpet. Mocny wiatr przejmował do szpiku kości i Ta-lia postanowiła przejść ze szkolnej izby do

8 , Gisou? - Nie. Nazajutrz, gdy przyszedłem go zwolnić, gromada małp siedziała mu na głowie, ramionach i plecach. Ciągnęły go za włosy, **kąsały** w uszy i wpychały palce w nozdrza, oczy i usta. Nerwowe tiki wykrzywiały mu twarz tak zabawnie, że wybuchnąłem śmiechem. - Jesteś zadowolony, Gisou

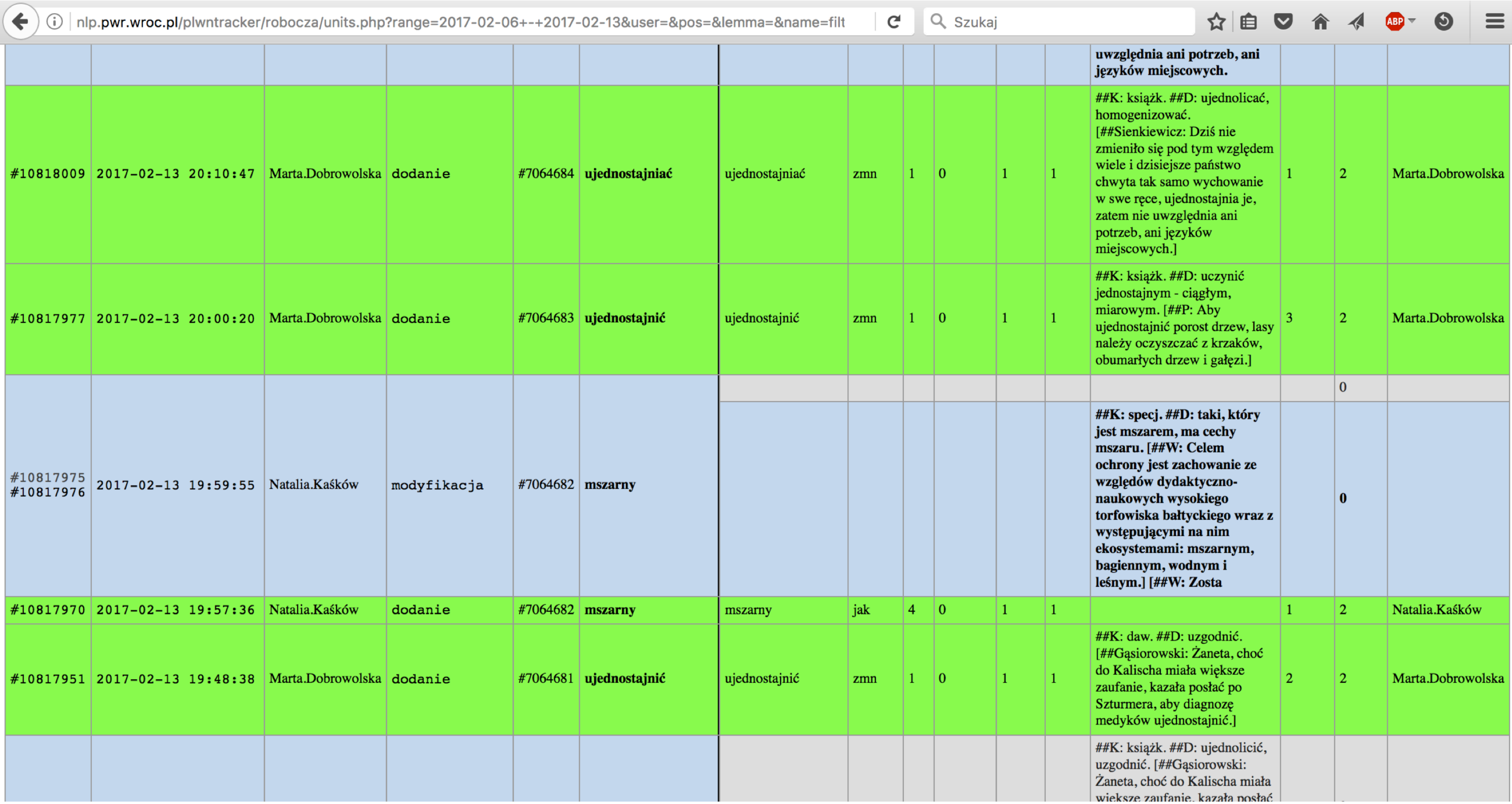
9 Artaq zdążył już minąć druida i kłębowisko demonów; jego lśniące ciało kołysało się równo, gdy biegł ku otwartej równinie. Kilka ciemnych kształtów rzuciło się na nich, **kąsając** ostrymi zębami nogi koni. Artaq nie zwalniał. Kopnął nogą jednego z demonów i odrzucił go daleko od siebie. Pozostałe zwolniły kroku. Wil pochylił się nisko,

10 głos. — Dziś w nocy nastąpił masowy wylęg bąblowca ryjkowatego. Wezwano nas trochę za późno. Część niebezpiecznych owadów przedostała się już do sanatorium i kąsają. — **Kąsają**? Nie zauważyłem. — Ukąszenia bąblowców są bezbolesne, dopiero po godzinie zaczyna się nieznośne swędzenie, wyskakują na ciele bąble, a potem następuje najgorsze stadium:

Corpus-based Wordnet Development

6. Wordnet editing supported by tools
    - Semantic exploration of corpora
        - Corpus concordancers
        - *LexCSD - usage examples – primary source for adjectives and verbs*
        - *Measure of Semantic Relatedness*
        - *WordnetWeaver*
    - Consulting traditional linguistic resources
        - dictionaries, encyclopaedias (including Wikipedia), lexicons…
        - linguists’ intuition, guidelines and consulting within plWordNet team
7. Linguistic work management
    - System for group work (Redmine system): tasks assignment, team communication etc.
    - plWordNet `Big Brother’ – a web-based system for monitoring and verifying work
    - Verification and coordination: linguists plus coordinators

plWordNet `Big Brother’


nlp.pwr.wroc.pl/plwntracker/robocza/units.php?range=2017-02-06+-+2017-02-13&user=&pos=&lemma=&name=filt
Szukaj

												uwzględnia ani potrzeb, ani języków miejscowych.			
#10818009	2017-02-13 20:10:47	Marta.Dobrowolska	dodanie	#7064684	**ujednostajniać**	ujednostajniać	zmn	1	0	1	1	##K: książk. ##D: ujednolicać, homogenizować. [##Sienkiewicz: Dziś nie zmieniło się pod tym względem wiele i dzisiejsze państwo chwyta tak samo wychowanie w swe ręce, ujednostajnia je, zatem nie uwzględnia ani potrzeb, ani języków miejscowych.]	1	2	Marta.Dobrowolska
#10817977	2017-02-13 20:00:20	Marta.Dobrowolska	dodanie	#7064683	**ujednostajnić**	ujednostajnić	zmn	1	0	1	1	##K: książk. ##D: uczynić jednostajnym - ciągłym, miarowym. [##P: Aby ujednostajnić porost drzew, lasy należy oczyszczać z krzaków, obumarłych drzew i gałęzi.]	3	2	Marta.Dobrowolska
#10817975 #10817976	2017-02-13 19:59:55	Natalia.Kaśków	modyfikacja	#7064682	**mszarny**									0	
												##K: specj. ##D: taki, który jest mszarem, ma cechy mszaru. [##W: Celem ochrony jest zachowanie ze względów dydaktyczno-naukowych wysokiego torfowiska bałtyckiego wraz z występującymi na nim ekosystemami: mszarnym, bagiennym, wodnym i leśnym.] [##W: Zosta		**0**	
#10817970	2017-02-13 19:57:36	Natalia.Kaśków	dodanie	#7064682	**mszarny**	mszarny	jak	4	0	1	1		1	2	Natalia.Kaśków
#10817951	2017-02-13 19:48:38	Marta.Dobrowolska	dodanie	#7064681	**ujednostajnić**	ujednostajnić	zmn	1	0	1	1	##K: daw. ##D: uzgodnić. [##Gąsiorowski: Żaneta, choć do Kalischa miała większe zaufanie, kazała posłać po Szturmera, aby diagnozę medyków ujednostajnić.]	2	2	Marta.Dobrowolska
												##K: książk. ##D: ujednolicić, uzgodnić. [##Gąsiorowski: Żaneta, choć do Kalischa miała większe zaufanie, kazała posłać			

Synset relations

- Hypernymy/hyponymy
  - defined for all parts of speech
  - also for verbs
  - adjectives and adverbs – limited but surprisingly numerous
- Inter-register synonymy
  - nouns, verbs, adjectives and adverbs

  ≈ synonymy across different stylistic registers
  - links stylistically marked lexical units with their unmarked counterparts
  - e.g. *samochód* ‘a car’ – *fura* ‘a car (slang)’

Substitution tests

Condition:

Stylistic register of *Y* must be not lower in the register hierarchy than register of *X*.

Testing expressions:

If she/it is *X*, then she/it must be *Y*

If she/it is *Y*, then she/it need not be *X*

If she/it is not *Y*, then she/it cannot be *X*

Substitution tests

Condition:

Both: *ocean* ‘ocean’ and *zbiornik wodny* ‘water basin’ are of the general stylistic register.

Testing expressions:

If she/it is *oceanem* ‘ocean’, then she/it must be *zbiornikiem wodnym* ‘water basin’

If she/it is *zbiornikiem wodnym* ‘water basin’, then she/it need not be *oceanem* ‘ocean’

If she/it is not *zbiornikiem wodnym* ‘water basin’, then she/it cannot be *oceanem* ‘ocean’

Noun Lexical Relations

- Contrast
  - Complementary antonymy
  - Proper antonymy
  - Converseness
- Cross-PoS Synonymy (N-V, N-Adj)
- Feminity
- Markedness
  - Young being
  - Deminutive
  - Augmentativeness
- Feature bearer
- Role
  - Agens
  - Instrument
  - Result
  - Place
  - Patient
  - Time
  - Result with unexpressed predicate
  - Place with unexpressed predicate
- Derivation

Lexical unit relations

- **Antonymy**
  - all parts of speech
  - *complementarity* – polar pairs of LUs with opposite and mutually exclusive meanings
  - *gradable opposition* – non-exhaustive oppositions
- **Converseness**
  - nouns and verbs
  - mutually opposite roles assigned to the arguments
  - for nouns:

If A is X (Prep) B, then B is Y (Prep) A

e.g., *If A is a husband of B, then B is a wife of A*

Synset Relations

- Hyponymy and hyperonymy
- Backward relations
  - *presuposition* (V-V,N,A,Adv) - close to logical presupposition
    - *żywy* 8 'alive' ←pres.- *umrzeć* 1 'to die'
  - *preceding* (V-V,N,A,Adv) - represents a possibility that one situation happens before the other one
    - *siedzieć* 1 'to sit', *stać* 3 'to stand' ←prec- *położyć się* 1 'to have laid down'
- Co-occurrence of two situations
  - *meronymy* (V-V_{imp}) and holonymy (V-V_{imp}) (not automatically reverse) - a situation is an element of a larger, more general situation, necessary simultaneous co-occurrence of two situations
    - meronymy: *przełykać* 'to swallow' is an integral part of situation *jeść* 'to eat'
    - holonymy: *jeść* 'to eat' is a typical situation including *przełykać* 'to swallow'

Synset Relations

- Beginning of a situation
  - *inchoativity* (V-V_{imp},N), where the first verb represents an initial phase of a situation represented by the second element
    - *zakochać się* 1 'to fall in love' → *kochać* 1 'to love'
- Resulting in a situation
  - *processuality* (V-N,A,Adv) – 'to become or to be becoming
    - *zmieniać się* 1 `to be changing itself/yourself' = to be becoming '*inny* 1 `different'
  - *causality* (V-V,N,A,Adv) – 'to cause'
    - *zablokować 2* 'to lock' → *blokada* 4 'lock'
- State (V-N,Adj,Adv) – being in some state
  - *jaśnieć* 1 'to shine_{imp}' means 'to be bright [*jasny* 8]' or 'to be brightly [jasno 8]'

Synset Relations

- Multiplicativity
  - *Iterativity* (V_{imp}-V) – repetition of some state or activity
    - *grywać* '~to play a little from time to time' → *grać* 'to play'
  - *Distributivity* (V_{perf}-V_{perf}) – performing an activity by many subjects or on many objects
    - *nakupić* 'to buy many things' → *kupić* 'to buy$_{perf}$'
- Inter-register synonymy V-V
  - LUs are close in meaning but have incompatible stylistic registers
  - *pieprzyć* [vulgar] '~to speak (nonsense)' → *mówić* 'to speak'

Non-constitutive Synset Relations

- Manner [V-Adv]
  - describes a verb by a link to an adverb describing a manner in which the activity is performed
  - substitution test
    - `If smn/smth has *X*, it means that he/it has *Z Y* [Adv]', where *X* is a hyponym of *Z*'
  - *popracować* 'to work$_{perf}$ a **little**' → ***trochę*** [Adv] 'little [Adv]'
- Circumstance [V-N]
  - describes a verb by referring to an adverbial realised by a simple prepositional phrase and its noun head
  - *dopłynąć* '~to swim$_{perf}$ to some point/place' -circum.→ *brzeg* `a bank'
- Subject [V-N] and object [V-N]
  - *subject: muczeć* 'to moo'→ *krowa* 'a cow'
  - object: *obuć* 'to put on **shoe**'→ ***but*** 'a shoe'

plWordNet content

	synsets	lemmas	LUs	avs
GermaNet	101,371	119,231	131,814	–
Princeton WordNet 3.1	117,659	155,593	206,978	1.74
enWordNet 1.0	+7841 125,500	+10119 165,712	+11633 218,611	1.74
plWordNet 4.0 emo	**222,137**	**191,447**	**288,074**	**1.32**

- LUs – lexical units (= senses)
- avs – average synset size

PLWORDNET
SŁOWOSIEĆ

plWordNet content

- 53 different relation types (107 when counting subtypes)
  - including many relations linking lexical units of different PoS
- Semantic domains (*lexicographer files* from WordNet)
- Semantic verb classes – constitutive features, supporting defining the relation structure
- Stylistic labels (11 in total)

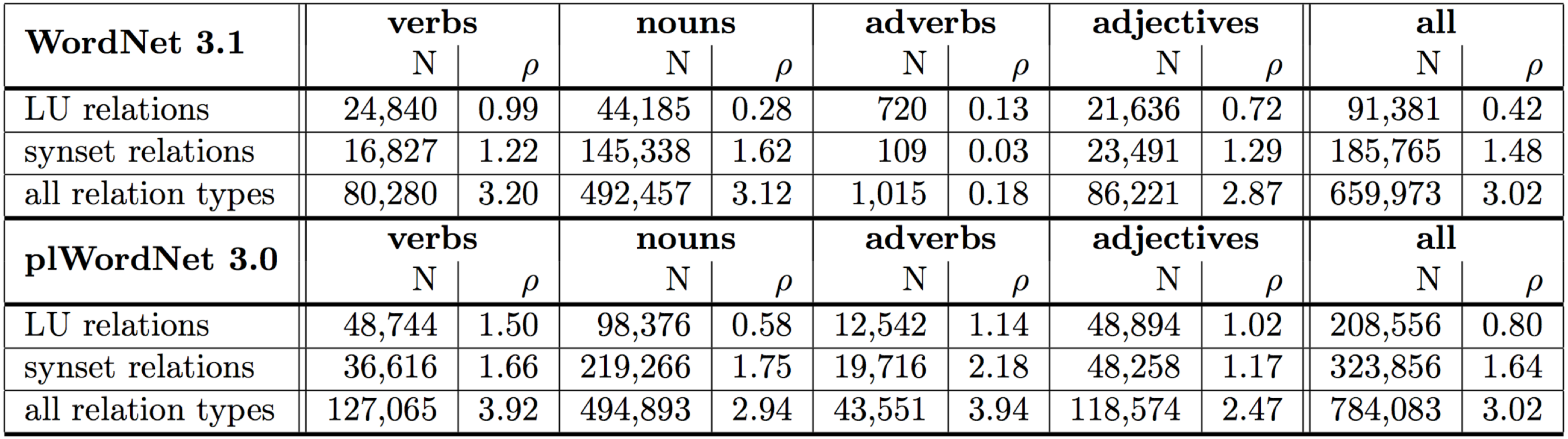
Description layer	Instances
lexico-semantic relations	~716K
glosses	>163K
usage examples	>73K(+ ~36K emotive)
links to Wikipedia	~55K
sentiment annotation	>86K

plWordNet emotive annotation

- Basic emotions
  - joy, trust, fear, surprise, sadness, disgust, anger, anticipation (Plutchik, 1980)
- Fundamental human values (Puzynina, 1992)
  - (positive) utility, another's good, truth, knowledge, beauty, happiness
  - (all negative) futility, harm, ignorance, error, ugliness, misfortune
- Sentiment polarity
  - +strong +weak, neutral, -weak, -strong
- Usage examples for positive and negative annotations
- Details: cf (Zaśko-Zielińska et. al, 2015) from RANLP' 2015 and (Zaśko-Zielińska & Piasecki, 2018) from GWC'2018

plWordNet comparison

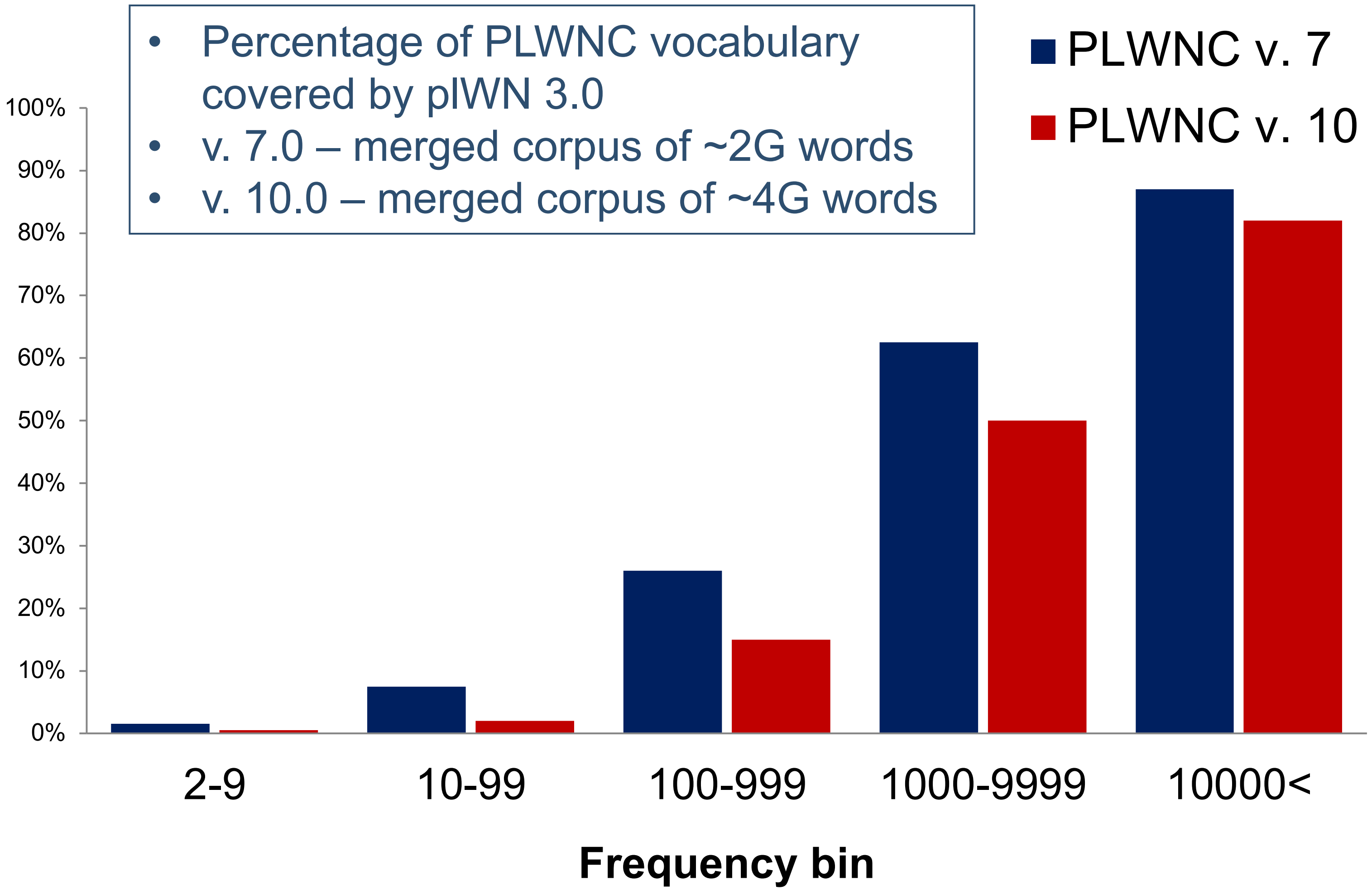
Network volume and density



WordNet 3.1	**verbs**		**nouns**		**adverbs**		**adjectives**		**all**	
	N	ρ	N	ρ	N	ρ	N	ρ	N	ρ
LU relations	24,840	0.99	44,185	0.28	720	0.13	21,636	0.72	91,381	0.42
synset relations	16,827	1.22	145,338	1.62	109	0.03	23,491	1.29	185,765	1.48
all relation types	80,280	3.20	492,457	3.12	1,015	0.18	86,221	2.87	659,973	3.02
plWordNet 3.0	**verbs**		**nouns**		**adverbs**		**adjectives**		**all**	
	N	ρ	N	ρ	N	ρ	N	ρ	N	ρ
LU relations	48,744	1.50	98,376	0.58	12,542	1.14	48,894	1.02	208,556	0.80
synset relations	36,616	1.66	219,266	1.75	19,716	2.18	48,258	1.17	323,856	1.64
all relation types	127,065	3.92	494,893	2.94	43,551	3.94	118,574	2.47	784,083	3.02

ρ is the relation density measured either for LUs, or synsets, or for all relation types

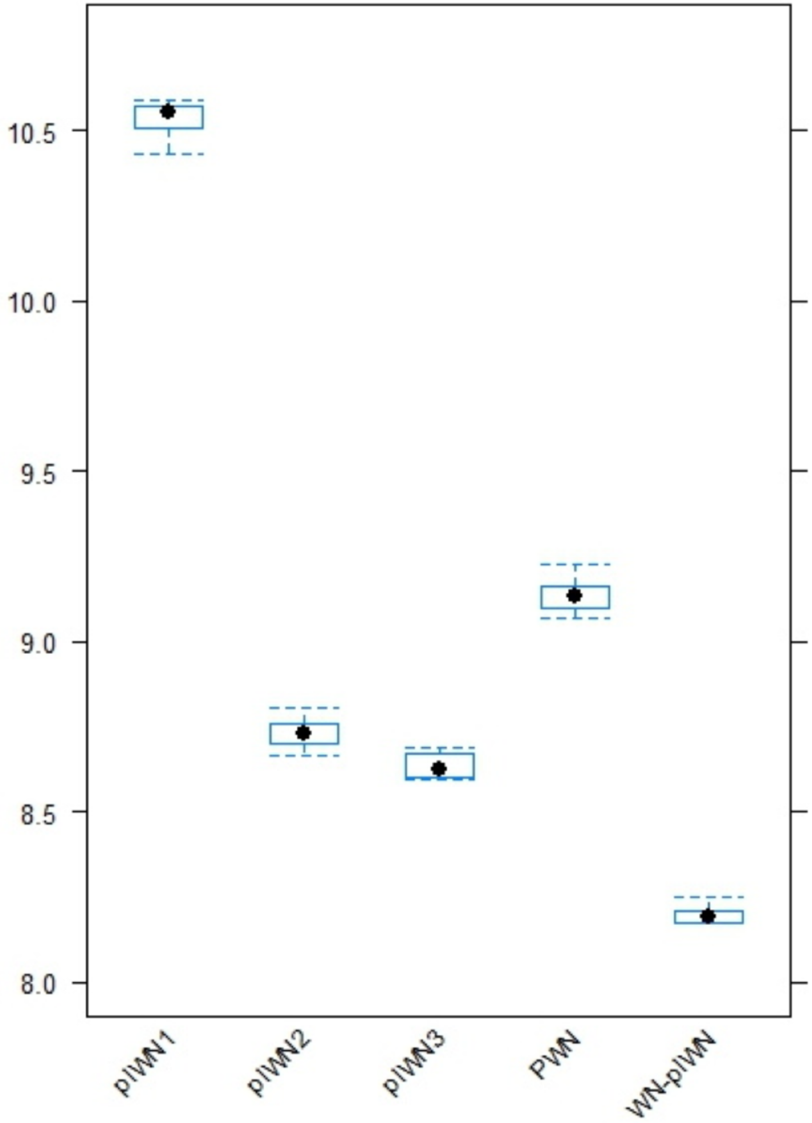
plWordNet coverage



- Percentage of PLWNC vocabulary covered by plWN 3.0
- v. 7.0 – merged corpus of ~2G words
- v. 10.0 – merged corpus of ~4G words

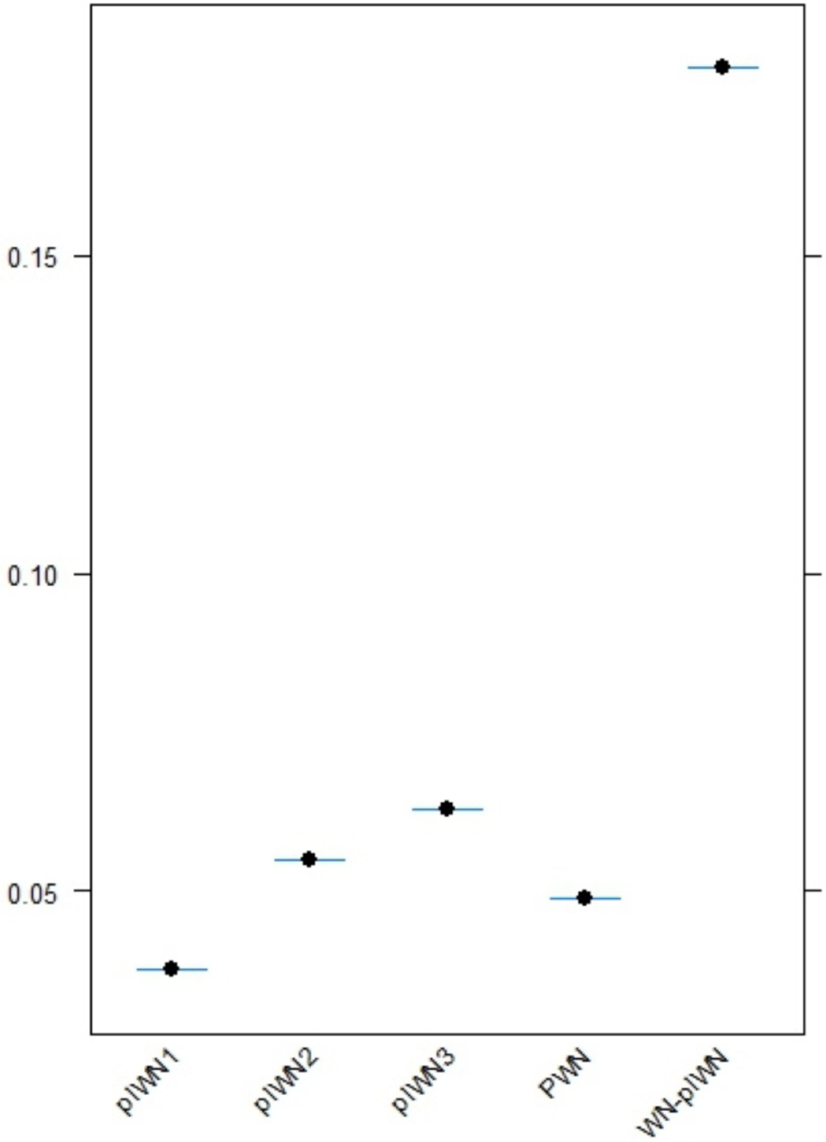
plWordNet as a small world

- Compared graphs
  - plWN1-plWN3 – different version of plWordNet
  - PWN – WordNet 3.1
  - WN-plWN – PWN and plWN 3.0 combined by mapping
- Average path length
  - expected: short average path length
  - average path length for the random graph: 11



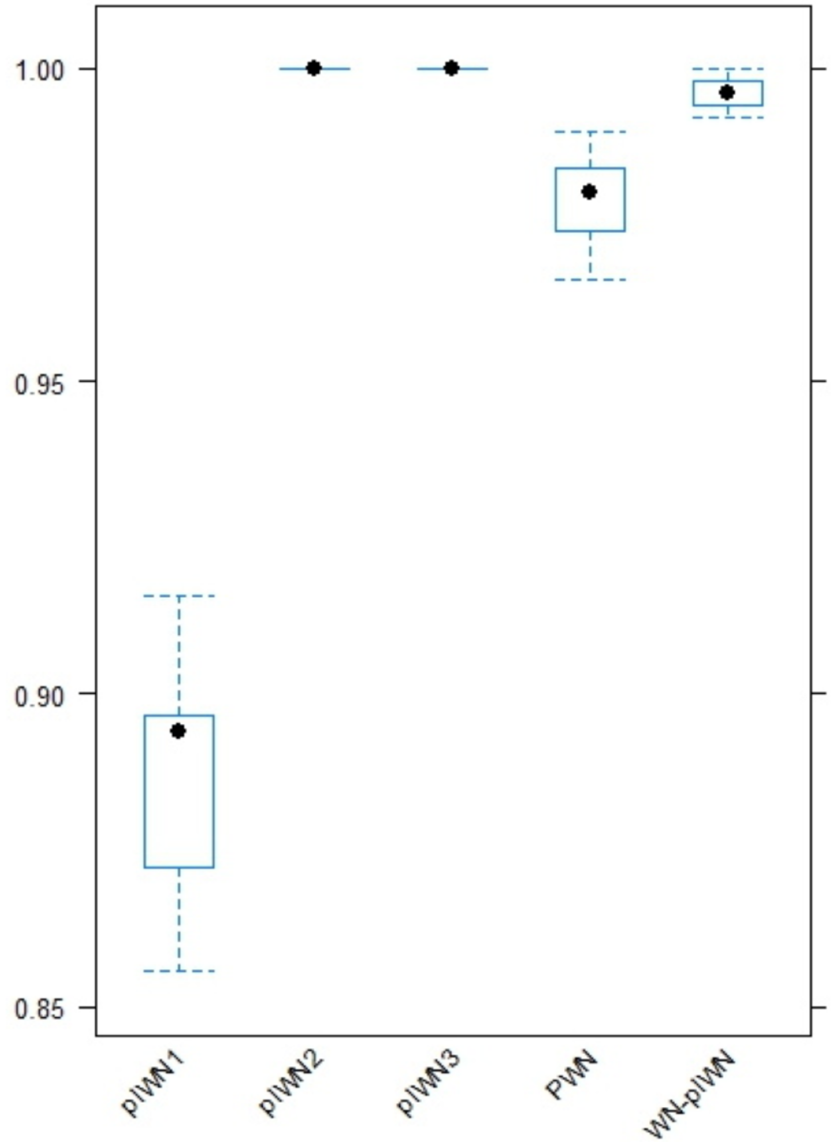
plWordNet as a small world

- Global clustering coefficient
  - higher - denser



plWordNet as a small world

- Connectivity
  - how often a path can be established between two synsets randomly chosen



plWordNet mapping

- Recognise the sense of a source synset by:
  - its position in the network structure,
  - existing relations, commentaries (glosses),
  - comparison to other synsets containing the given lemma
- Search for candidates for a target synset:
  - intuitions, automatic prompting and dictionaries
- Verify candidates:
  - by comparing hypernymy and hyponymy structures
  - by exploring existing inter-lingual relations;
  - by comparing definitions, commentaries; dictionaries
- Link the source synset with the target synset

plWordNet 4.0 emo applications

- Wide coverage inspires a lot of applications
- plWordNet is a pivotal element of a system of language and knowledge resources
- An anchor to Linked Open Data via mapping to WordNet
- Monolingual and bilingual dictionary
  - Web-based: http://plwordnet.pwr.edu.pl
  - Android application
  - WordnetLoom for visual, graph-based browsing
  - included in a very large and popular Polish multilingual Web dictionary Ling
- WordTies (Pedersen et al., 2012), Open Multilingual WordNet (Bond and Foster, 2013)

plWordNet 4.0 emo applications

- Numerous research applications, for instance
  - Classification of gestures based on the verb categorisation in plWordNet (Lis and Navarretta, 2014)
  - Referred to in the resource for textual entailment (Przepiórkowski, 2015)
  - Language correction
  - Relation extraction
  - Text classification
  - Open Domain Question Answering
  - A quasi-ontology in document structure recognition
- An exceptional case is the practical use of plWordNet during the medical treatment of aphasia

plWordNet 3.0 emo applications

- A large number of declared applications:
- Education (at different levels) including Polish language teaching,
- Building dictionaries, extraction of synonyms and semantically related words, detection of loanwords,
- Cross-linguistic study on phonesthemes, classification of metaphorical expressions,
- Corpus studies, grammar development, comparative and contrastive studies,
- Language recognition, parsing disambiguation, semantic analysis of text, document similarity measures, semantic indexing of documents, semantic information retrieval,
- Recommendation systems, construction of chatbots and dialogue systems,
- Plagiarism detection,
- Translation evaluation, data visualisation, research on complex networks and ontologies, …

Conclusions

- Corpus-based wordnet development methods allows for good coverage of language data and close relation to the language use
- Minimal Commitment Principle wordnet models results in a relational semantic dictionary
- plWordNet is an example of a wordnet which is a relational semantic dictionary and *vice versa*

Thank you very much for your attention!

www.clarin-pl.eu

CLARIN-PL
Common Language Resources and Technology Infrastructure

Supported by the Polish Ministry of Science and Higher Education [CLARIN-PL]

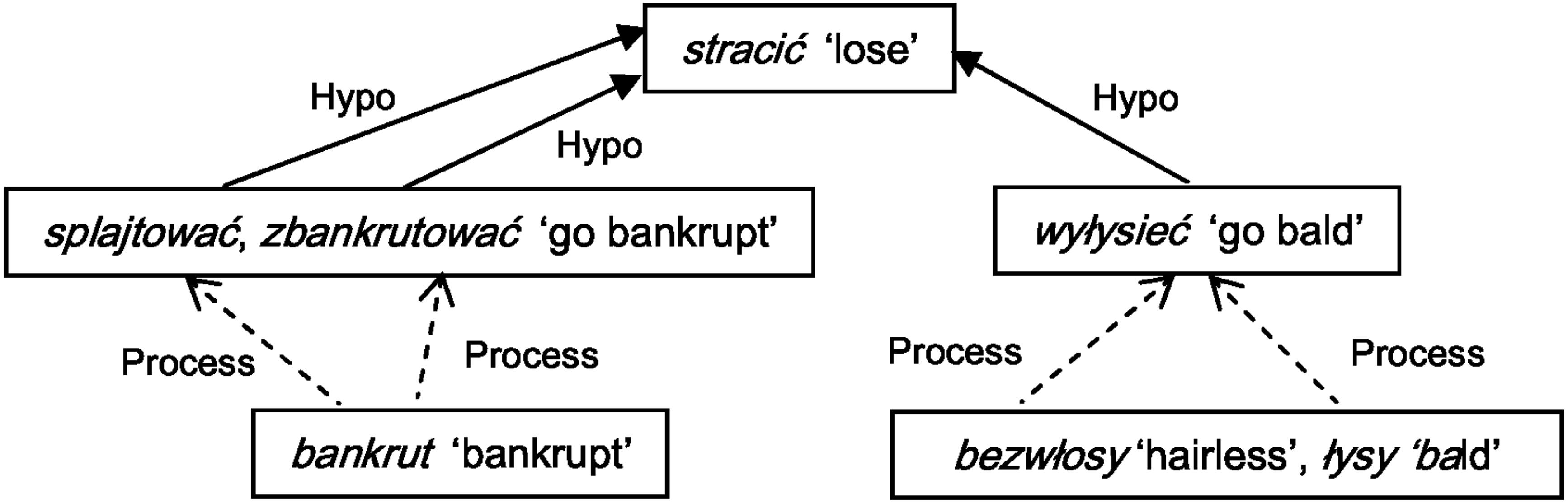
Bibliography

- B. Broda M. Marcińczuk, Maziarz Radziszewski Wardyński, Adam (2012). KPWr: Towards a Free Corpus of Polish. In Khalid Choukri et al. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012, ISBN: 978-2-9517408-7-7.
- Rudnicka Ewa, Maziarz Marek M, Piasecki Maciej, Szpakowicz Stanisław (2012) A stategy of mapping Polish WordNet onto Pinceton WordNet. In 24th International Conference on Computational Linguistics : proceedings of COLING 2012, 8-15 December 2012, Mumbai, India : posters. Vol. 3. Mumbai : The COLING 2012 Organizing Committee, 2012. s. 1039-1048.
- Kurc, R.; Piasecki, M. & Broda, B. (2012) Constraint Based Description of Polish Multiword Expressions. In Calzolari, N.; Choukri, K.; Declerck, T.; Dogan, M. U.; Maegaard, B.; Mariani, J.; Odijk, J. & Piperidis, S. (Eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), 2408-2413.
- Woliński, Marcin. (2014) Morfeusz Reloaded. In Nicoletta Calzolari (Conference Chair) Khalid Choukri, Thierry Declerck Hrafn Loftsson Bente Maegaard Joseph Mariani Asuncion Moreno Jan Odijk Stelios Piperidis (Ed.): Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp. 1106–1111, European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, ISBN: 978-2-9517408-8-4.
  http://nlp.ipipan.waw.pl/Bib/wol:14.pdf

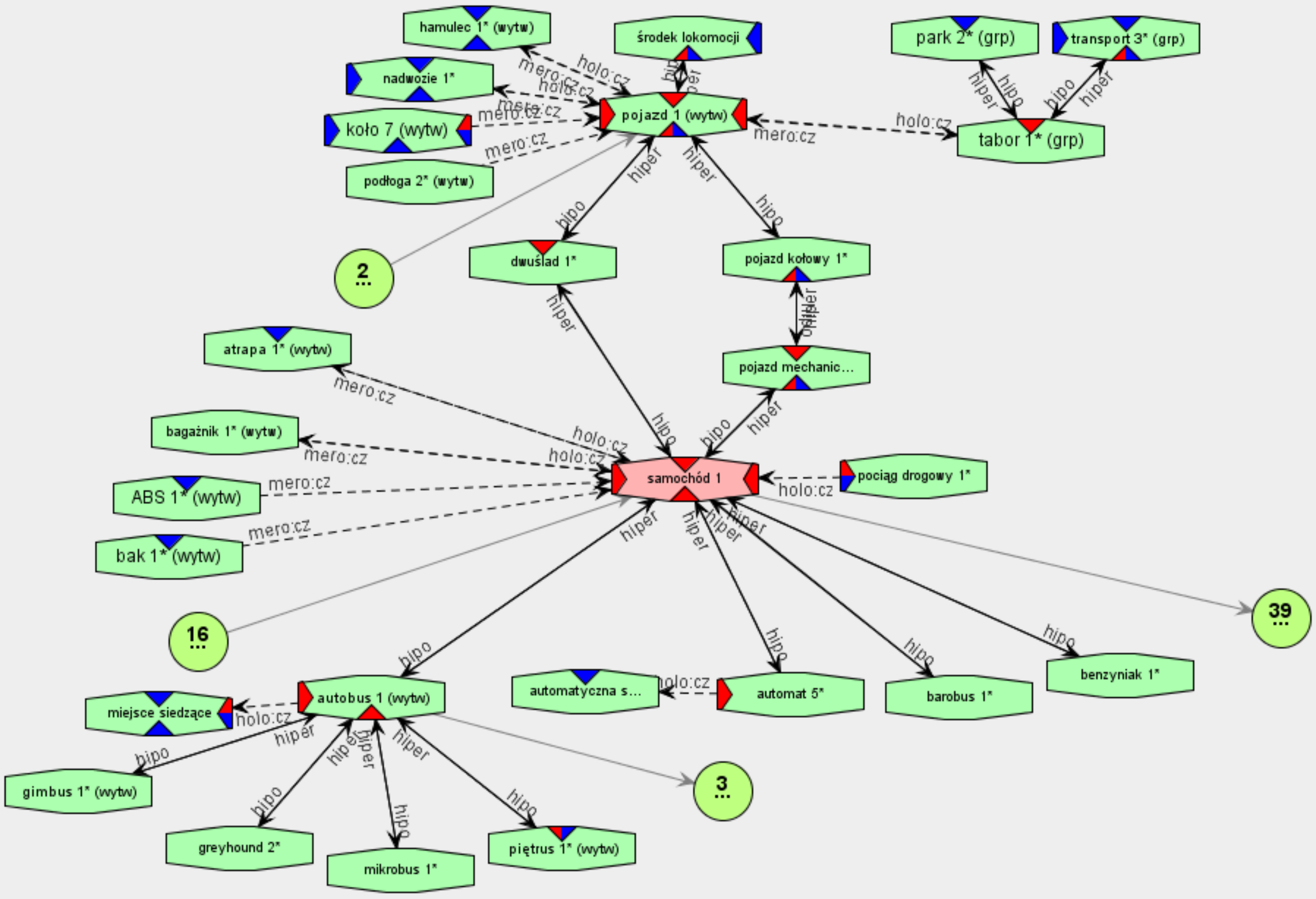
Bibliography

- Przepiórkowski, Adam; Hajnicz, Elżbieta; Patejuk, Agnieszka; Woliński, Marcin; Skwarski, Filip; Świdziński, Marek Walenty (2014) Towards a comprehensive valence dictionary of Polish. In Choukri, Khalid et al. (Ed.): Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp. 2785–2792, European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, ISBN: 978-2-9517408-8-4.
- Maziarz, M.; Piasecki, M. & Szpakowicz, S. (2013)The chicken-and-egg problem in wordnet design: synonymy, synsets and constitutive relations Langauge Resources and Evaluation, 2013, 47, 769-796, http://link.springer.com/article/10.1007/s10579-012-9209-9 *(open access)*
- Maziarz, Marek and Piasecki, Maciej and Rudnicka, Ewa and Szpakowicz, Stan and Kędzia, Paweł (2016) plWordNet 3.0 -- a Comprehensive Lexical-Semantic Resource. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. The COLING 2016 Organizing Committee, pp. 2259—2268, Osaka, Japan http://aclweb.org/anthology/C16-1213

Constitutive relations: example



plWordNet (Słowosieć)



Paintball: knowledge sources

- Knowledge sources $K_1, \ldots K_s$ extracted by different methods from the corpus
- $K_i = \{ <l_n, l_j, w>:$

  l_n – a new word,
  (not in the wordnet)
  l_j – a wordnet word

  w – local weight (for the pair) }
- $weight(K_i) \in (0,1]$ – global weight
  (for the knowledge source)

(Piasecki et al., 2013a)

Paintball algorithm

- Input: a wordnet, a new word and a set of Knowledge Sources
- Output: a set of subgraps – attachment areas – with one synset marked in each
- Idea
  - each knowledge source expresses some error level
  - knowledge source triples are not precise in pointing to particular synsets
  - hits covers regions
  - spreading activation helps to analyse and combine the delivered information

Evaluation: method

- Evaluation by reconstruction
  - a word sample is removed from the wordnet
  - *Paintball* is applied to reattach the words
- Data collected
  - histogram of path lengths between suggested synsets and the original positions in a wordnet
  - paths of up to 5 links, including hyper/hyponymy links with at most one final meronymic were considered

Evaluation: method

- Criteria
  - **closest** path: attachment proposition that is closest to the original location
  - **strongest** suggestion: top scored
  - **all** suggestions

Evaluation: experiment setup

- Wikipedia corpus, including almost 1 billion words
- Word sample
  - corpus frequency threshold for words: 200
  - words that have at least 3 hypernymy links to the top synset
  - 1064 test words selected
  - margin of error 3% and 95% confidence level
  - frequent words ≥ 1000
  - infrequent words ≤ 999

Evaluation: baseline

- Baseline: *Probabilistic Wordnet Expansion* (Snow, Jurafsky, & Ng, 2006)
  - lack of procedure for setting the values of parameters
  - selected experimentally:
    - minimal probability of evidence: 0.1,
    - inverse odds of the prior: $k = 4$,
    - maximum size of the cousins neighbourhood: $(m, n) \leq (3,3)$,
    - maximum links in hypernym graph: 10
    - penalization factor: = 0.9

Results: straight path strategy

Method			Hit distance								
			0	1	2	3	4	5	6	**[0-2]**	Σ
PWE	Rare	C	3.7	21.7	16.2	9.6	6.9	3.4	0.1	41.6	**61.5**
		S	0.5	5.9	9.7	10.9	8.9	4.5	0.5	16.1	**40.9**
		A	0.8	4.9	5.0	4.5	3.8	2.0	0.4	10.7	**21.5**
	Freq.	C	0.8	14.8	24.2	21.0	15.1	5.5	0.2	39.8	**81.6**
		S	0.1	2.7	9.4	16.1	15.7	13.2	0.8	12.2	**58.0**
		A	0.2	3.2	7.0	10.0	9.8	7.3	0.5	10.4	**38.0**
PB	Rare	C	**9.2**	21.7	12.6	6.7	4.2	1.0	0.6	**43.5**	56.1
		S	**4.8**	**13.1**	10.0	6.5	3.4	1.2	0.4	**27.9**	39.4
		A	**2.9**	**6.9**	4.8	3.5	2.2	1.0	0.2	**14.6**	**21.5**
	Freq.	C	**6.3**	**20.5**	15.0	11.9	6.7	2.6	0.5	**41.8**	63.3
		S	**1.9**	**9.1**	8.4	8.1	4.8	1.9	0.3	**19.4**	34.7
		A	**1.4**	**4.9**	4.4	4.4	3.1	1.6	0.2	**10.7**	20.0

Results: folded path strategy

Method			Hit distance					
			0	1	2	3	4	Σ
PWE	Rare	C	3.7	21.7	18.4	11.8	2.5	58.2
		S	0.5	5.9	10.7	12.6	2.3	32.0
		A	0.8	4.9	6.6	6.9	1.5	20.7
	Freq.	C	0.8	14.8	25.2	22.9	4.0	67.7
		S	0.1	2.7	9.6	17.0	3.4	32.8
		A	0.2	3.2	7.9	12.2	2.9	26.4
PB	Rare	C	9.2	21.7	21.9	10.7	1.9	**65.5**
		S	4.8	13.1	15.3	13.1	1.5	**47.9**
		A	2.9	6.9	14.7	13.2	1.7	**39.4**
	Freq.	C	6.3	20.5	20.7	18.6	2.8	**68.8**
		S	1.9	9.1	11.5	13.5	3.1	**39.2**
		A	1.4	4.9	8.4	11.6	2.3	**28.5**

Results: coverage

- For the straight path strategy
- Coverage for words
  - PWE: propositions for 100% of words (freq. 100%)
  - Paintball: 63.15% of words (freq. 91.93%)
- Recall for senses
  - PWE: 44.79% (freq. 43.93%)
  - Paintball : 24.66% (freq. 26.62%)

Results: example

- PWE suggestions for *feminism*

  {*abstraction*, *abstract entity*},
  {*entity*},
  {*communication*},
  {*group*, *grouping*},
  {*state*}

- *Paintball* suggestions:

  {*causal agent*, *cause*, *causal agency*},
  {*change*},
  {*political orientation*, *ideology*, *political theory*},
  {*discipline*, *subject*, *subject area*, *subject field*, *field*, *field of study*, *study*, *bailiwick*},
  {*topic*, *subject*, *issue*, *matter*}

Corpus-based Development of Lexico-semantic Resoures

- MWELexicon
  - extraction of collocations from corpora
  - verification of lexicalisation supported by decision trees suggested by ML methods
- NELexicon 2.0
  - extraction from the Web
  - NER applied to large corpora
  - verification supported by Wikipedia info-boxes
- Walenty
  - A large syntactic-semantic valency lexicon
  - based on examples extracted from the National Corpus of Polish, deep parser pre-processing, treebank annotation and careful manual editing and verification

Synset relations

- Holonymy/meronymy
  - between nouns, but also verbs
- Meronymy
  - subtypes: *part*, *element of a collection*, *place*, *portion*, *substance*

    {*iskra 1*, *skierka 1*, *skra 1*, *iskierka 1* 'spark'} —m.part→ {*ogień 1*}

    {*kula 4* 'bullet', *nabój 2 'cartridge'*} —m.e.coll.→ {*amunicja 1* 'ammunition'}

    {*drewno 1* 'wood', *drzewo 1* ~'timber'} —m.subst.→ {*stolarka 2* 'woodwork'}

    {*termin 1* 'fixed date', *data 1* 'date'} —m.place.→ {*czas 1* 'time'}

    {*blacha 1* 'sheet metal'} —m.port.→{*metal 2* 'metal'}
  - taxonomic unit
    - e.g. *kotowate* (Felidae) – *kotokształtne* (Feliformia)
  - special sub-typea for verbs: accompaning situation, e.g. *chrapać* 1 to snor - *spać* 1 to sleep
    - Holonymy is not automatically reverse, e.g. . *szprycha* 'spoke'—m.part→ *koło* 'wheel'

Synset relations

- Type/instance
  - Proper Names linked to common nouns
- Inhabitant
  - based on the derivational relation,
    but expanded to synsets
  - at least two synset members must be derivationally associated
  - e.g., *Japończyk* ‘Japanese’ – *Japonia* ‘Japan’

    {Trojańczyk, Trojanin} ‘Troy citizen’ – {Troja, Ilion} ‘Troy’

    Troj-ańczyk, *Troj-anin* < *Troja*

    Troj-ańczyk, *Troj-anin* < *Ilion* (but they pass tests

Derivational relations

- **Cross-categorial synonymy**

  ~ `NEAR’ relations in EuroWordNet

  `transpositional’ or `syntactic’ derivation

  POS shift without any significant semantic change

  pis-anie `writing (gerund)’ < pisać `write’ (N-V)

  pisz-ący `writing (part.)’ < *pisać* (Part.-V)

  czerwon-ość `redness’ < *czerwony* `red’ (N-Adj)

Derivational relations

- **Feature|State bearer** (N-Adj)

  ślepi-ec `a blind man’ < *ślep-y* `blind’

  starz-ec `an old man’ < *star*-y `old’

  Feature|State is an inverse relation

- **Femininity** (N-N)

  feminine derivatives from masculine bases

  pisar-ka `~writeress’ < *pisarz* `writer’

  kot-ka `female-cat’ < *kot* `cat’

Derivational relations

- **Markedness** (N-N)
  - **diminutives** `small’ , `tiny’ , `nice’

    dom-ek `small or nice house’ < *dom* `house’

    książecz-ka `tiny or nice book’ < *książ-ka*
  - **augmentatives** `big’ , `large’ , `awful’

    ptasz-ysko `big or awful bird’ < *ptak* `bird’

    noch-al `big or awful nose’ < *nos* `nose’
  - **young being** `offspring child of an animal’

    koci-ę `kitten’ < *kot* `cat’

Lexical unit relations

- Role (examples)
  - Signalled derivationally, but with definite semantics
  - Agens
    - *biegacz* `runner’ – *biegać* `to run’
  - Instrument
    - *wiertarka* `driller’ – *wiercić* `to drill’
  - Result
    - *układ* `configuration’ - *układać* `~to configure’
  - Time
    - *świt* `dawn’ – *świtać* `to dawn’

enWordNet 1.0

- Motivations
  - I-Synonymy is more useful than I-Hyponymy, …
- Goal: to exploit the existing I-Hyponymy links to extend WordNet’s coverage
- Result
  - enWordNet 1.0 – an extended version of WordNet 3.1
  - focus given to translations of the plWordNet leaf synsets:
    - equivalents whose lemmas were not present in WordNet
    - no equivalents;
    - equivalents whose lemmas were already present in WordNet
  - verification with English corpora and dictionaries
  - other missing lexical units from hypernymy branches added
- **7 841 new English synsets, 11 633 LUs, 10 119 lemmas**

Walenty – a Large Valency Dictionary for Polish

- A large syntactic-semantic valency dictionary for Polish
- Sub-dictionaries
  - Verbs: >12 000 lemmas
  - Nouns: ~2 500 lemmas
  - Adjectives: ~950 lemmas
  - Adverbs: ~200 lemmas
- Two inter-connected levels of schema description
  - Morpho-syntactic
  - Semantic
- Rich phraseological information
- Usage examples for schema realisations
- http://zil.ipipan.waw.pl/Walenty; http://walenty.clarin-pl.eu/

Walenty – a Large Valency Dictionary for Polish

- Example of syntactic schema
- *adresować* `to address' (_, ,imperf) `sth to sth/someone'

Schema	verified [1]		
Function	Subj	Obj	
Phrase type	np(str)	np(str)	prepnp(*do*,gen)

- where
  - _ - any value of negation
  - np – Noun Phrase
  - prepnp – Prepositional Phrase
  - str – structural case,
    obj: acc/gen (in the case of negation)
    subj: nom/acc (in the case of modification by some
    numerals and quantifiers)

Walenty – a Large Valency Dictionary for Polish

- Example of syntactic schema
- *adresować* `to address’ (_, ,imperf) `letter/message/… to sth/someone’

Schema	**Verified [70]**		
Function	Subj	Obj	
Phrase type	np(str)	np(str)	prepnp(*na*,acc)

Walenty – a Large Valency Dictionary for Polish

- Example of semantic schema
- *adresować* 1 \`to address’ letter/message/… to sth/someone’

Frame	Verified [60416]		
Role	Recipient	Theme	Initiator
Selectional preferences	SUBJECTS	*propozycja* 1 \`proposition’	SUBJECTS

- PODMIOTY: {HUMAN, podmiot-3}

Walenty – a Large Valency Dictionary for Polish

- Example of semantic schema
- *Adresować* 2 \`to address' letter/message/… to sth/someone'

Frame	Verified [60416]		
Role	Theme, Goal	Theme, Source	Initiator
Selectional preferences	SUBJECTS	*przesyłka* 1 \`~mail'	SUBJECTS
	jednostka administracyjna 1 \`administrative unit'		
	miejscowość 1 \`an inhabitated place'		
	nazwa własna 1 'Proper Name'		

Walenty – Sets of Semantic Roles

Roles	Initiating Group	Accompaning Group	Closing Group
Basic	Initiator Stimulus	Theme Experiencer Factor Instrument	Recipient Result
Completing	Condition	Attribute Manner Measure Location Path Time Duration	Purpose
Attributes	Source	Foreground Background	Goal

Walenty – Linking Syntactic and Semantic Levels

- *Adresować* 2 `to address’ letter/message/… to sth/someone’

Frame	Verified [60416]		
Role	Theme, Goal	Theme, Source	Initiator
Selectional preferences	SUBJECTS	*przesyłka* 1 `~mail’	SUBJECTS
	jednostka administracyjna 1 `administrative unit’		
	miejscowość 1 `an inhabited place’		
	nazwa własna 1 ‘Proper Name’		

Schema	verified [1]		
Function	Subj	Obj	
Phrase type	np(str)	np(str)	prepnp (*do*,gen)

Schema	Verified [70]		
Function	Subj	Obj	
Phrase type	np(str)	np(str)	prepnp (*na*,acc)