Spokes PL

This page contains the documentation for the Spokes PL conversational search engine. Spokes PL currently gives access to a corpus of 2 319 291 words (247 580 utterances) of conversational Polish, which makes it a unique resource for scholars, researchers and engineers interested in the spoken register of Polish.

Please make sure you cite Spokes properly:

Pęzik, Piotr. “Spokes – a Search and Exploration Service for Conversational Corpus Data.” In Selected Papers from the CLARIN 2014 Conference, October 24-25, 2014, Soesterberg, The Netherlands, 99–109. Linköping Electronic Conference Proceedings. Linköping University Electronic Press, Linköpings universitet, 2015.

Here is a BibTeX record:

```
@inproceedings{pezik_spokes_2015,
    series = {Linköping {Electronic} {Conference} {Proceedings}},
    title = {Spokes — a search and exploration service for conversational
corpus data},
    copyright = {CC-BY-NC},
    isbn = {978-91-7685-954-4},
    url =
{http://www.ep.liu.se/ecp_article/index.en.aspx?issue=116;article=009},
    abstract = {Spokes is an online service for conversational corpus data
search and exploration, currently developed as part of CLARIN-PL — the
Polish CLARIN infrastructure. This paper describes the data sets currently
available through Spokes, the architecture of the service and the data and
metadata search functionality it provides to its users. We also introduce
some of the more experimental features which have been developed to
facilitate more advanced research on multimodal conversational corpora.},
    booktitle = {Selected {Papers} from {CLARIN} 2014},
    publisher = {Linköping University Electronic Press, Linköpings
universitet},
    author = {Pęzik, Piotr},
    year = {2015},
    pages = {99--109}
}
```

SlopeQ syntax

Spokes uses the SlopeQ 2 query syntax. The examples below are customized to show how the SlopeQ syntax can be used for searching the Polish conversational data sets we provide through Spokes. For practical reasons the number of examples illustrating each query in this presentation is very limited. However, a link to a page with all the results is given for each query.

Surface queries

This is the simplest type of queries. You input words in plain written form in the query box. The results are occurrences of the particular forms submitted in the query. Compare the query and a selection of results:

`mamo`

#	Left	Match	Right
1	ale to nie ale to jest naprawdę niepotrzebne	mamo	
2	tak	mamo	przyznam oczywiście
3	wiesz co	mamo	zadzwonię za jakieś dwadzieścia minut dobra
4	daj mi	mamo	keczupku
5	ale	mamo	ale to to jest już prawie czyste ta miska

Queries of this kind can include a sequence of terms (positions in the query). A series of words is put in query box and the results will show occurrences of the whole sequence, e.g.:

`wiesz co`

#	Left	Match	Right
1	a teraz	wiesz co	jest taka sprawa pan był załatwiać tam ten ten Leszek tak
2	latanie bardzo mi się podoba start lądowanie	wiesz co	prawda a później to tak wiesz
3		wiesz co	ja skończę ten
4	no	wiesz co	właśnie tak ale są nie podobni wiesz
5	no no bo	wiesz co	to nawet jest włóczka taka fajna do takich to musi być
6	nie wiem bo	wiesz co	tylko wystawy widziałem że przecenione rzeczy
7	no nie wiem właśnie co się stało z tą kasetą ze się	wiesz co	
8		wiesz co	że może nie o to chodzi że nie jest taka w naszym wieku
9	Kaśka	wiesz co	w tym są moje włosy nie bierz tego
10		wiesz co	ze tak od trzydziestu do pięćdziesięciu

This type of queries is useful with set phrases and some collocations.

Base form queries

These queries make use of the lexical annotation of the data in the Spokes corpus. The results are the occurrences of different forms of the given word. The query is written in triangular brackets as an equation “lemma=”. After the equation mark you put the base form of the word.

The base form of the word is: the **infinitive** for verbs, **nominative singular** for nouns (except for pluralia tantum), **nominative singular masculine** for adjectives.

`<lemma=palić>`

`palić**`

#	Left	Match	Right
1	to	pal	pierwsza o co ci chodzi
2	tutaj	palą	kominkiem
3	jak nie palą przecież teraz	palisz	i teraz idź na całe osiem godzin nie pal to bycie tam pokręciło
4	ale na jakiś czas ale jak ona codziennie	pali	to
5	no skumała nie no bo przecież może	palić	nie tam wszyscy palą
6	a paliła i babcia	paliła	i dziadek
7	mów dalej bo to kurwa ciekawe jest czemu ja wtedy zioła nie	paliłem	

This kind of query is obviously very useful in languages like Polish where nouns and verbs have numerous grammatical forms. It is a simple way to research the use of the word in all its forms.

Note: some grammatically possible forms of a searched word may not appear in the data.

Note: the results illustrate different meanings of the queried word – compare the first example of palić above with the others.

You may also put several base forms in the query box. The occurrences of combinations of various forms of the words found as a sequence will appear in the results, e.g.

`mieć** szansa**`

#	Left	Match	Right
1	Kaśka dopóki nie będziemy miały internetu to nie	ma szans	
2	to tak a bez matury to nie	ma szans	
3	dlatego ponieważ opisywali to wszystko że ten ma szanse ten	ma szanse	i i ten wiesz
4	przecież jeszcze z planem rozłożonym wiesz mam	mam szansę	zaj spojrzeć wiesz nikt mi nie wyskoczy
5	to	miał szansę	być typowany trzy miejsca w klasie do sz przez szkołę na studia
6	i trudno było a ja w zakładzie kupiłam bo	miałam szansę	
7	misiu nie	mają szansy	żadnej misiu o teraz będzie metalika

This is a very useful and productive type of queries for investigating particular collocations. The order of the terms in the above query is fixed (see below [link] for queries with any order of the elements).

It is also possible to put a surface form and a base form as two terms in one query. The example shows what forms of the pronoun ja follow the form słuchaj in the corpus:

`słuchaj ja**`

#	Left	Match	Right
1		słuchaj ja	pogadam z nim bo niewykluczone że on tam zna kogoś no
2		słuchaj mnie	czy ty masz może klucz

#	Left	Match	Right
3	a już ci przody robię ale ty wiesz jak ja siedzę do godziny dziesiątej siedzę i robię na drutach bo	słuchaj mi	się teraz zwaliło roboty
4	a to się tak rzadko zdarza żeby było w mieszkaniu takie miejsce na taki pawlacz tutaj było wszystko załadowane ten cały pokój był zagracony rzeczami ja mówię	słuchaj ja	nie wiem gdzie my to pochowamy no i pawlacz rozwiązał sytuacje wszystko się zmieściło
5	ale ja nie	słuchaj ja	nie ściągałam urządzeniem
6	bo to jest właśnie to że	słuchaj ja	to w ogóle właśnie jak gadam i tak dalej robię z siebie taką niesamowitą językoznawczynię i jeżeli chodzi o akcenty i o metodykę i o to że ona jest taka bardzo mocno British i tak dalej

Operators

Alternative

This operator is represented by the pipe sign “|”. The words separated by the operator are variants of the query term. The results are occurrences of all of them, e.g.:

`tu|tutaj`

#	Left	Match	Right
1	Marek kiedyś przejechał pamiętasz nie bo jak byłam mała mama	tutaj	nam przywiozła takiego
2	no i mówię ci ciotka dulczyła tutaj ciotka	tu	dulczyła żeby Gośka nakręciła nie
3	bo teraz pasjoniści nasi też tu mają że	tutaj	też się uczą podobno y w seminarium tutaj a kto wykłada to nie wiem
4	nie wiem czy to jest seminarium czy tylko jest taki po prostu że oni	tu	są mieszkają może
5	ale chwileczkę bo późni polecisz a	tu	nic może bym ze trzy kartofle starła i placki

It is possible to combine more than two options in a query, e.g.:

`wokół|dookoła|dokoła`

#	Left	Match	Right
1	tak	wokół	rury
2	to jest dosyć skomplikowana taka taka dziewczyna zresztą bardzo ładna i dlatego wykorzystuje wszystkich	dookoła	
3	obeszliśmy to jeziorko	dokoła	
4	odpowiedni robi	wokół	tego a wszystkim zależy urzędasom żeby wiesz tam tego a rzeczywistość jest niestety mało ekologiczna

#	Left	Match	Right
5	i to był taki dom co się	dookoła	chodziło i była taka sypialnia Piotrek z Rafałem byli mali

The variants in the query can also be base forms. In that case, occurrences of all forms of all the words will appear in the results, e.g.

`facet**|koleś**`

#	Left	Match	Right
1	miśku to o czyją wolność on walczył o wolność	facetów	
2	taki obraz taki obraz	kolesia	wyłania się z szałasa opowieści
3	on jest taki śmieszny ten	koleś	
4	ja współczuję temu	facetowi	
5	ogłoszenie	facet	dał ze wynajmie mieszkanie
6	no bo później jakichś dwóch	kolesi	było którzy maja chyba dziekankę wzięli i się nie będą bronić no i jeszcze jedną laskę co ten co się broni we wrześniu
7	no i jak zaczęła się szafa zgadaliśmy się że	faceci	robią zabudowy okna no to też zrobimy
8	Przemek mi opowiadał że jakiś	koleś	podrabiał pięciozłotówki nie i nie wiedzą teraz co z nim zrobić bo okazało się za za te monety płacił dzieciom za posiłki w szkole no i normalnie za podrabianie pieniędzy jest dożywocie ponoć

As the examples show, the most natural application of such queries is with synonymic words.

It is also possible to use the alternative operator for one of the terms in a longer query. In the example below, the results are occurrences of the sequences:

`bardzo|strasznie dużo`

#	Left	Match	Right
1	o na tym podwórzu to było tyle dzieci i było tyle tylu tych lokatorów nie było tak znowuż	bardzo dużo	
2	tych banków to	bardzo dużo	w Kaliszu jest czy w Łodzi też tyle jest
3	francuski nie dużo jest	bardzo dużo	
4	teraz dopiero odkryłem że nawet zdaje się wiesz zanim się wyrwiesz z gdzieś za miasto to musisz przejechać tutaj kurde w tym smrodzie dymie tymi ulicami	strasznie dużo	
5		strasznie dużo	szkód im ten pies robi taki
6		strasznie dużo	książek macie

Slop factor

This important functionality allows to search for a discontinuous string of words. The query specifies how many words may intervene between the terms of the query. This parameter is known as the slop factor and it is now set in the search menu, e.g.:

`ta kobita (Slop factor = 1)`

#	Left	Match	Right
1	myślałem wiesz już znaczy wiesz	ta kobita	przez nich właśnie trafiła do tego depeesu przez tą żonę tego faceta no i jego samego bo tam wiesz była granda taka
2	a ona poszła i mówi tak mnie molestuje mówi	ta Andrzeja kobita	bo ma tyle pieniędzy kredytu
3	i jest wiesz no ale mówię i ta i	ta jego kobita	też jest niegłupia chyba i taka wiesz sensowna i tam wiesz
4	tak jak przy tej poprzedniej zusowskiej kontroli ja mam nadzieję że tutaj wiesz specjalnie chociaż	ta kobita	jest taka ostra

Note: the provided number is the maximum number of intervening words. Strings with fewer or no intervening words will also appear in the results.

Slop factor queries can have query terms other than surface forms. For example, you can also use base forms, e.g.

`jechać** tam (Slop factor = 2)`

#	Left	Match	Right
1	czy	jechać sprawdzić tam	na tego Limanowskiego
2	to bez sensu my wypijmy tą kawę i	jedźmy się tam	przywitać i po drodze po chleb wejdziemy
3	w sensie żeby nie czekać do wakacji tylko już teraz na przykład na wakacje	jedziesz sobie tam	i żywego języka słuchasz i już wiesz
4	czyli nie tutaj gdzieś tam jak się na to Skrzyczne	jedzie tylko tam	dalej gdzieś jeszcze
5	jak	jadą tam	twój to tam nie ma siedzeń ani nic to tam żeby były siedzenia to
6	a samochodem nie	jechałem pociągiem raz tam	jechałem

Slop factor with relaxed order

These queries allow intervening words up to the specified number and the query terms may appear in any order. This parameter is now set in the search menu, e.g.:

`jest sprawa (Slop factor = 2, Unordered terms)`

#	Left	Match	Right
1	a teraz wiesz co	jest taka sprawa	pan był załatwiać tam ten ten Leszek tak
2	albo	sprawa jakaś jest	w toku czy coś myślę że oni wtedy też nie są w domu dziecka tylko właśnie w tym pogotowiu
3	trzy miesiące ostanie w tej szkole rodzenia ale to	jest bardzo fajna sprawa	naprawdę wiesz

#	Left	Match	Right
4	to tak sobie pomyślałam co ty pierdolisz no jak nie swoje sprawy że to w sumie	jest jej sprawa	jest moją przyjaciółką no to
5	jej nie pasuje tak to nie chodziło o to że	jest sprawa	nie do załatwienia i tak mnie to wkurwia na każdym kroku i właśnie chodzę taka sfrustrowana
6	bo to ciekawa	sprawa jest	nie
7	inna	sprawa to jest	taka że oni są dosyć dziecinni oni

The relaxed order in pure form is available with the number 0, i.e. with no intervening words:

```
no nie (Slop factor = 0, Unordered terms)
```

#	Left	Match	Right
1	że masz już z głowy	nie no	to super
2		nie no	jeszcze nabiorę sobie sama
3	tak samo właściwie na ciepło można by podać z tym ryżem no	nie no	co
4	co	no nie	jest to mało na pewno nie jest to mało ale
5	znaczy tak planujemy na długość	no nie	
6	i raczej tak	no nie	powiem grzeczni są

Relaxed order may be combined with other functionalities, for example with that of alternative:

```
dziś|dzisiaj jest (Slop factor = 0, Unordered terms)
```

#	Left	Match	Right
1	a co	dzisiaj jest	
2		dziś jest	pięknie najładniejszy dzień dzisiaj
3	no może	jest dzisiaj	za gorąco ej są takie podkładki chłodzące nie coś takiego jest
4	weź go tyknij może się ruszy właśnie nie wiem co	jest dziś	tak właśnie wiesz bo nawet jeszcze wczoraj tak nie robił a dzisiaj no kurde

In the next example, the base form query, multiple term input, and slop factor with relaxed order are used together. As a result, the searched sequences contain various forms of the verb czekać, the two query terms appear in different orders and there may be one intervening word between them:

```
<lemma=czekać> na (Slop factor = 1)
```

#	Left	Match	Right
1	ale już wczoraj wczoraj	czekałam na	film bo zawsze modę na sukces oglądam
2		czekam na	kontrakt
3	ból był straszny już mi tam podkładali różne rzeczy bo ja tydzień	na operację czekałam	bo miałam tu krwiak straszny musiał zejść

#	Left	Match	Right
4		czekając na	przykład do lekarza po lipne zwolnienie na egzamin bo jesteś bardzo chory a tak naprawdę zapiłaś się dzień wcześniej i masz po prostu kaca
5	no bo my wtedy babcia jak weszłyśmy do tej rodziny oni się pytaja chce pani dzisiaj to my tak i	czekała babcia na	zabiegi
6	no jemu byłoby raźniej myślę miałby	na kogo czekać	miałby opiekę jakby coś tam gorzej to chyba z takim dziadkiem

Negation

This operator excludes specified variants of query terms from the results. Consequently, it must be combined with query types that produce variation in the results. Negation is marked by a pipe sign with an exclamation mark "|!", which is to be read as "but not". The example shows how it is used with a base form query. The specified form of the word is excluded from the results:

```
<lemma=prosić>|!proszę
```

#	Left	Match	Right
1	mówiła że ona tylko młodych	prosi	a starych nie chce widzieć
2	mogę cię	prosić	
3	sam powinien zrezygnować a nie jeszcze żeby go	prosili	
4	jak ja go się	prosiłam	żeby zlazł
5		proszą	się o robotę
6	no no	prosze	was przecież co się je z waciki nasączone sokiem owocowym i to na tydzień masz jedzenie z głowy

Queries of this kind may be used to exclude word forms that have special properties or particularly high frequencies and thus skew the data like the form *proszę* excluded in the example.

Regex queries

Queries of this type make use of special symbols and quantifiers. Each query is a formula describing a whole set of possible strings of signs (words, sequences of words). The results are occurrences of all predefined strings found in the data.

Wild card and quantifiers

A full stop "." is a wild card, it stands for any sign.

A plus "+" is a quantifier: the preceding sign can appear one or more times.

An asterisk "*" is another quantifier: the preceding sign can appear zero or more times.

These symbols can be used directly with standard signs, but the most fruitful use in queries is to combine the wild card with one of the quantifiers.

“.+” means that in this part of the query any sign or sequence of signs may appear.

“.*” means that in this part of the query any sign or sequence of signs or nothing may appear.

Note the difference between the quantifiers. If you use the plus, the preceding symbol (which may be any symbol if you use the wild card) needs to appear at least once in each item found. With the asterisk it may not appear at all.

Compare the examples:

`tam.+`

#	Left	Match	Right
1	bieg takiej a takiej rzeki poprzez budowę takiej a takiej	tamy	czy tam przeniesienie koryta nie
2	Oleńka	tamta	miała
3	ty żeś już jedną żeśta chałupę w prezencie dostali co drugą chcesz dostać mówię	tamtą	ojciec tobie zostawił
4	syn gospodarza	tamtego	tego domu
5		tamto	mi puściło listki na razie
6	jedno jabłko musiałem wykroić bo już zaczęło gnić tylko tyle się zostało z	tamtych	jabłek które chcesz
7	rodzeństwem bo	tamci	po kryjomu podobno dzwonią tam czasami przed tamtymi rodzicami ale

The above query searches any word starting with “tam” but not plain *tam* because at least one sign must appear after these three letters. To include plain *tam* in the results, the other quantifier has to be used:

`tam.*`

#	Left	Match	Right
1	nie teraz były kiedyś	tam	
2	tak żeśmy się popisywały ja tobą a ona	tamtą	
3	Ewelina teraz ma kolegę Karola i jak opowiada o tym Karolu to ciągle sobie nie mogę jeszcze przestawić się na Karola	tamtego	
4	a co coś	tam	a może argo
5	a później idę z kościoła i ta leci za mną i Lolcia Lolcia bo ona tu ma syna w	tamtym	bloku pod dwójką
6	ta	tama	w Włocławku to samo lichutka już tyż ją chcą remontować
7	poszedł na dwór to te pchły gdzieś	tam	by poodskakiwały pozlatywały by z niego

The use of wild card without quantifiers allows for crossword-like queries which yield word forms of a set number of letters, containing set letters at certain positions, e.g.:

`k..t.`

#	Left	Match	Right
1	ale mieliście	karty	bankomatowe tam można normalnie tak płacić kartą
2	teraz jak masz	kartę	kochana to za kartą jeszcze musisz mieć konto oczywiście
3	teraz jak masz kartę kochana to za kartą jeszcze musisz mieć	konto	oczywiście
4	no i masz no i	klatę	robisz jeszcze w dipsach nie na poręczach
5	są też klity takie wiesz stare gdzie płacisz wiesz sto złotych czy tam a są te nowe bloki gdzieś kurcze na Powiślu czy gdzieś	kwoty	kosmiczne
6	to jest sprytny wynalazek że ona ma trzy	knoty	

You can use the wild card and a quantifier several times in one query, e.g.

```
t.* bab.*
```

#	Left	Match	Right
1	ale może to z zamianą	tego babci	mieszkania tak ale nie wiem na czym to polega czy można to zamienić czy nie można
2	ale jak ona miała na imię to nie pamiętam taka już wiesz co no koło czterdziestki	też babeczka	też bardzo sympatyczna
3	słuchaj taka była	ta baba	no ja myślałam że wyleci
4	on wyjechał do Anglii jego córka ta mała	ta babcia	
5	no tak ta wyższa byłaby lepsza do	takiej babki	
6	ale jeszcze gorsi od	tych bab	są nie prawdziwi niedzielni kierowcy
7	a jak to było z	tym babunem	
8	niedaleko naszego domu	taka babeczka	nam sprzedawała to wiesz zawsze jak tam byłyśmy to

The wild card and quantifiers can also be combined with other functionalities. You can use them in a base form query. The results are occurrences of all forms of all words allowed by the formula in the query, e.g.

```
<lemma=p.+biec>
```

#	Left	Match	Right
1	w łazience to zimno jest w łazience	pobiegła	gdzie jesteś w sypialni no tu masz cieplej aj i tam słyszę zaraz znalazła
2	wiecie co szczur zdobył się na heroiczny czyn	podbiegł	tam pod pod blok i przy ścianie biegł ten ptak tam wiesz
3	mi się to już wolę iść spać i rano	przebiec	dziesięć kilometrów
4	no coś nam	przebiegło	chyba nie
5	i na pewno	przybiegnie	na pewno przybiegnie zawsze przybiegają Adama zobaczyliśmy po jakiejś pół godzinie

In the next example two formulae using wild card and quantifiers are variants in a query with the alternative operator.

szykow.+|przygotow.+

#	Left	Match	Right
1	w tym akurat temacie to jest dosyć dużo bo muszę cały koncert	przygotować	na pierwszy dyplom
2	no przecież ja się	przygotowałam	na in-class się można przygotować
3	bo oni mają tam potencjał ogromny przecież ta sztuka rosyjska ten balet ta opera ta architektura przecież oni maja	przygotowaną	kadrę taką jak nie wiem
4	taki do	przygotowania	samodzielnego
5	w sensie żeby oni	przygotowywali	prezentacje
6	no przynajmniej nie straciłaś czasu na	przygotowywanie	się
7	ale dużo stałam no bo ten obiad tez	szykowaliśmy	
8	gdzieś tam słyszałem że będzie lepsze stanowisko	szykowane	
9		szykowanie	się na imprezy i właśnie stroje tak facet i

In the example below, the query with wild card and quantifier is restricted by the negation operator:

zna.*|!znaczy

#	Left	Match	Right
1	tak no trzeba by mieć przepisy trzeba by pogadać z kimś kto się na tym	zna	no bo wiesz zero jakiejkolwiek reakcji no i nic nie możesz zrobić
2	ale patrz jak już ty	znasz	to cała Polska go zna
3	i jeszcze chciał żebyśmy napisali zdania z tymi czasownikami że wiemy co	znaczą	nie a ja w ogóle zapomniałam o tym kole i się w ogóle nie nauczyłam
4	idzie raczek nieboraczek jak ugryzie będzie	znaczek	
5	to ma duże	znaczenie	no a ona się wykuje po prostu recytować i potem to pisze w słowo w słowo
6	no zaraz ci ich	znajdę	
7	sami	znajomi	tutaj jest mama
8	jakiś tam terenowy samochód się	znalazł	gdzieś tam Mariusza zawiózł do lekarza
9	aaa ja tam go nie	znałam	nawet
10	jak	znam	życie

Grammatical queries

The format of the grammatical annotation

The Spokes corpus makes use of the tagset developed for the NKJP. The grammatical distinctions made in the tagset are fairly detailed and not necessarily obvious at first. This presentation deals only with the basic issues relevant for queries in the Spokes corpus.

A comprehensive description of the tagset and the categories used in the annotation can be found in the NKJP handbook (Ł.Szałkiewicz and A. Przepiórkowski 2012. “Anotacja morfoskładniowa.” In: Przepiórkowski et al. *Narodowy Korpus Języka Polskiego*. Warszawa: PWN) on pages 62-67, further

relevant details are discussed on pages 69-81 [link]. The help page for the Poliqarp search engine also provides information on the categories and values in the tagset .

In order to submit successful grammatical queries, you must know how the grammatical information in the corpus is organised.

This is an example of a short sentence fragment (*podszedłem do gościa*) from Spokes as it is tagged and stored in the database. As can be seen, the surface form, the grammatical information and the base form are three key components of the annotation.


```
[
  [
    {
      "orth": "Podszedł",
      "lexes": [
        {
          "CTag": "praet:sg:m1:perf",
          "base": "podejść",
          "disamb": true
        }
      ]
    },
    {
      "orth": "em",
      "lexes": [
        {
          "CTag": "aglt:sg:pri:imperf:wok",
          "base": "być",
          "disamb": true
        }
      ]
    },
    {
      "orth": "do",
      "lexes": [
        {
          "CTag": "prep:gen",
          "base": "do",
          "disamb": true
        }
      ]
    },
    {
      "orth": "gościa",
      "lexes": [
        {
          "CTag": "subst:sg:gen:m1",
          "base": "gość",
          "disamb": true
        }
      ]
    },
    {
      "orth": ".",
      "lexes": [
        {
          "CTag": "interp",
          "base": ".",
          "disamb": true
        }
      ]
    }
  ]
]
```


Grammatical queries search for specified sequences of signs in the grammatical part of the annotation. It is essential to know what is the order in which grammatical categories appear there. Quite naturally, different parts of speech have different sets of grammatical categories.

The following is the grammatical tagging of the noun form from the example (*gościa*). Pay attention to the order of the items of grammatical information:

“subst:sg:gen:m1”

part-of-speech:number:case:gender

Note: the labels in the tagset are fine-grained, e.g. “subst” covers most nouns, but not the “depreciative” forms like chłopy or komuchy, “m1” is masculine personal gender.

Note: **all noun forms** can be searched with the alias “noun” as the part-of-speech label instead of more detailed labels.

Note: grammatical words with syntactic properties of nouns (e.g. *coś*, *kto*) are classified as nouns.

Here is an example of the grammatical tagging of an adjective form (*ewidentnym*):

“adj:sg:inst:n:pos”

part-of-speech:number:case:gender:degree

Note: grammatical words with syntactic properties of adjectives (e.g. *ten*, *taki*) are classified as adjectives.

The grammatical annotation of verbs is most complex. Particular sets of verb forms have different categories associated with them. We will start with the past tense form (*podszedłem*) from the above example:

“praet:sg:m1:perf”

part-of-speech:number:gender:aspect

This is the annotation of a present tense form (*mówię*):

“fin:sg:pri:imperf”

part-of-speech:number:person:aspect

This is the annotation of an infinitive (*odpisywać*):

“inf:imperf”

part-of-speech:aspect

Note: because the relevant grammatical categories vary so much between different classes of verb forms, particular sets of forms (only three of them are shown above) are technically treated as different parts of speech in the tagset.

Note: **all verb forms** can be searched with the alias “verb” as the part-of-speech label instead of more detailed labels.

Note: verbal nouns (like *czytanie*, *spanie*) are tagged as gerunds, they are covered by the alias “noun”, but not by the alias “verb”.

Simple grammatical queries

Grammatical queries are put in triangular brackets and have the form of an equation with the label “pos=” (standing for “part of speech”) used in all of them. What is searched for is the whole string of grammatical information (see examples in the preceding section [link]). Since most grammatical queries are concerned with selected features only, they must make use of wild cards and quantifiers (see Regex queries [link]). The wild cards stand for the categories whose values are not specified in the given query.

Note: an obvious exception are words that do not have any grammatical information specified (many adverbs, conjunctions, interjections, etc.). However, not all uninflected words belong here (e.g. infinitives have aspect marked, prepositions have case marked, etc.)

In order to find all noun forms in plural, we must specify the part-of-speech information and the information about number and mark the rest of the grammatical annotation as unspecified using the wild card and quantifier:

`<pos=noun:subst:pl:.*>`

#	Left	Match	Right
1	może mi wreszcie ku ktoś kupi mówi kozaczki bo przecież nie będę zimą w	adidasach	chodziła
2	wziła ci	alimenty	bezczelnie i mówi że tego
3	nie ma kolejki	Austriacy	powiedzieli że nie uruchomią jej
4	myślała że ja zadzwonię dzisiaj z	życzeniami	no bo wiedziała że mam dzieci to nie mogę przyjść i chciała mnie jak zadzwonię
5	u znajomych teraz byliśmy taki gostek z biura turystycznego i mówi że we Włoszech właśnie jego znajomi byli na nartach on zresztą chyba w tym roku zacznie organizować	wyjazdy	też w do w Dolomity
6	znaczy n na niektórych	odcinkach	jeszcze można se łańcuchy wypożyczyć jak ich nie będziesz używał to ci oddadzą z powrotem pieniądze nie
7		perfumami	skropiła nas

The next query yields occurrences of noun forms in the instrumental case. The formula must mark unspecified information twice – for number (placed between part of speech and case) and gender (after case) – see the format of the tagging of nouns in the previous section.

`<pos=noun:subst:.+:inst:.+>`

#	Left	Match	Right
1	może praca oświatowa z	Aborygenami	
2	albo	bokiem	tak
3	tak albo że że ona nadrabia twarzą że są lepsze figury tylko jak ona	buzią	bardzo nadrabia
4	z tymi	dzieciakami	to ci mówię
5	przyjechał chłopaszek do niego	hondą	mu się zapomniało i za pięćdziesiąt tysięcy przyjechał nie
6	nie będziesz miał siły do wracania z	powrotem	
7	i ta kuratorka franca panie za	przeproszeniem	do mnie łazi i nie wim po co
8	no a tutaj ta książka to się	wszystkim	podobała

#	Left	Match	Right
9	bo to wiesz	zimą	wszyscy poubierani tak wielowarstwowo no
10	po	czym	poznać że słoń był w lodówce

The above queries makes use of the part-of-speech label “subst”, which accounts for most nouns, but not the “depreciative” forms like *chłopy* or *komuchy*. There is a general label (alias) “noun” which covers depreciative nouns, non-depreciative nouns as well as verbal nouns in the results.

A version of the previous query with the alias is shown below:

```
<pos=noun:.+:inst:.+>
```

#	Left	Match	Right
1	tak no trzeba by mieć przepisy trzeba by pogadać z	kimś	kto się na tym zna no bo wiesz zero jakiejkolwiek reakcji no i nic nie możesz zrobić
2	słuchaj może też za jego	plecami	to załatwili
3	y my jechaliśmy drugą	trasą	przez Rumunię Bułgarię
4	w ogóle nie przepadam za	wodą	szczerze mówiąc
5	tak moim	zdaniem	
6	zawsze tak z	gotowaniem	mówiłaś
7	BASIA przed	wyjściem	sobie powiedzmy że wychodzi o 21 wszystkie transakcje do 20 sobie już odznaczyła
8	to jest bez sensu z tym	spaniem	no ale jak my mamy o czwartej dopiero jechać to co ja mam robić znowu tak jeszcze druga trzecia to jeszcze pół biedy

As can be seen, there examples with verbal nouns lacking in the previous set of results. There are no examples of depreciative nouns because these are distinguished in the nominative case and not in the instrumental.

The next example shows a query for verbal forms in the present tense and the plural number.

```
<pos=verb:fin:pl:.*>
```

#	Left	Match	Right
1	gdzie się Jula schowała jest chowamy się	bawimy	się w chowanego
2	kasę	biją	jak nie wiem i mówi do niego naprzeciwko tak siedzieliśmy jak my przy stole
3	to jak	chcecie	lodziku
4	ale wtedy widać że że że wiesz	idą	seryjnie recepty że
5	ja Karolina właśnie	musimy	iść do do biura podróży jutro idziemy się zapytać ja Karolina Samanta Roksana i Aśka
6	a zbiórkę to	macie	tu i tu i o tej godzinie ja mówię no tak mamy nie i wiesz tak popatrzyłam
7	twierdzą że jest to w ogóle fa fajnie ekstra a okazuje się że no robią tak przez to	są	są podobni do do tych ludzi którzy tam przychodzą
8	różnie bo to wiesz ludzie usłyszą babcia wiesz usłyszy leki	tanieją	o osiemnaście procent

The query uses the detailed part-of-speech label “fin”, which marks non-past forms of verbs. The alias “verb” can be used, it covers past tense forms, non-past tense forms, impersonal past, infinitive, imperative, and all the participles.

Note: the tense distinction in verbs is shown by part-of-speech labels only. If the desired query is supposed to specify tense, the detailed part-of-speech labels are to be used. If tense is irrelevant for the query, the alias “verb” is probably the best label to use.

The part-of-speech label can be left unspecified like other parts of the grammatical annotation. In the next query, all occurrences of plural genitive forms are found:

```
<pos=.+:pl:gen:.+>
```

#	Left	Match	Right
1	bo tam nie ma tych takich ekonomicznych pewnie	barier	
2	bo tam nie ma tych takich	ekonomicznych	pewnie barier
3	a	ćwiczeń	nie zrobił wczoraj
4	ale ja myślę że w ciągu	dwóch	dni go skończę więc
5	przecież wszystko się robi dla swoich	dzieci	no czego się nie robi
6	tak potrafił jakoś mu się przypodobać że on zamiast iść do tych gdzie tamten go kierował do	tych	jakiś tam
7	no i tutaj też tak razem no razem jest mało jest	takich	małżeństw że tak razem przychodzą do kościoła ale oni byli i jeszcze jedno takie małżeństwo do
8	to były pomidory a ty nie nie byłaś Gosiu wtedy u	nas	
9	po	wszystkich	świętych to nie jest takie pilne

Let us stress again that in grammatical queries it is essential to get the order of information right. It is also essential to mark the irrelevant items of information as unspecified. Here are examples of queries that contain mistakes and yield no results:

<pos=noun> - no noun is marked just as noun, it has further grammatical information, which is not marked in the query <pos=noun:subst:inst:.+> - wrong order: case information does not come directly after part of speech information <pos=noun:subst:.+:f.+> - gender information comes last for nouns and is not followed by any signs

Grammatical queries combined with other functionalities

You may need to submit pure grammatical queries, but in many cases grammatical information will only be a part of the query you want to make. Here, we present combinations of grammatical queries with other possibilities of the query syntax of Spokes.

For example, you can combine base form query with grammatical query for a single term. The labels “lemma=” and “pos=” need to be taken in the same pair of brackets. Here, the feminine forms of the word niezły are searched for:

```
<lemma=niezły pos=.*:f:.*>
```

#	Left	Match	Right
1	ale ale po chwili już będzie więcej niż dwie jeszcze dwie o	niezła	mina uuu
2	no i to	niezła	
3	ale imprezę mieliście	niezłą	ej
4	no tam tam tam całkiem	niezłą	ekipę mieli
5	ale generalnie i tam wiesz ma całkiem chyba	niezłe	oceny

The query above does not specify the part of speech, but this is not necessary since niezły is unambiguously an adjective.

The next query yields occurrences of all singular forms of the verb zdać:

```
<lemma=zdać pos=verb:fin:sg:.*>
```

#	Left	Match	Right
1	sama przy tym nie będzie tylko tamta później nie wiem	zda	raport kurwa o mnie
2	tak te stare też już mają miały miedziane wiesz te te no i ma ten zapalnik pizoelektryczny trochę się obawiam czy	zda	egzamin
3	i przez ciebie nie	zdał	
4	a Karolina	zdała	
5	ja mówię aha nie mówię że no tak że	zdałam	sobie sprawę że później dopiero jak już wysłałam że to do ciebie doszło nie
6	dostałem się na studia	zdałem	egzaminy to co chciałem i te na które poszedłem to trzeba zaznaczyć że na tych na których byłem to zdawałem je
7	jest do zaliczenia na egzaminie najprościej kurwa męczyłam się męczyłam się tak z łaciną że nawet na w yyy na wrzesień miałam łacinę nie więc nie wiem jak ja to	zdam	nie po prostu jestem antytalencie jeżeli chodzi o łacinę
8	znaczy całkiem niezły no nie wiem B1 certyfikat miałem	zdany	na sehr gut

You can also combine the same two functionalities for two separate query terms. For this you need to take the appropriate labels in brackets separately. The example query shows occurrences of sequences of an adjective followed by any form of the word temat:

```
<pos=adj:.+> temat**
```

#	Left	Match	Right
1	stary ale zobacz jaki to jest chory w chuj temat jaki to jest w chuj	chory temat	chory temat nie pas koronowski w chuj szeroka no bo pas nie lotniczy y wojskowy
2	fajne były te teksty z angielskiego bo były na różne takie	ciekawe tematy	że można się było dowiedzieć różnych rzeczy no ale ten akurat był bardziej śmieszny niż jakiś
3	nie zupełnie nie	ten temat	miałem zupełnie powiedzieć na zupełnie inny temat tylko zapomniałem już w tym momencie

#	Left	Match	Right
4	czy coś konkretnego było na przykład że z	jakiegoś tematu	się mieliśmy przygotować
5	nie sieciowi znajomi to może być	poważniejszy temat	niż się wydaje wszystko
6	ci sami ludzie to samo do porozmawiania ten	sam temat	
7	no takie wiesz no tam ludzie podchodzą do	tego tematu	poważnie a tutaj nie
8	bardzo	trudny temat	wybrałam nie

This kind of query can be used to show collocations, the results of example query show adjectival collocations of the word *temat*.

Queries containing several terms can be further refined by using slop factor. The next example yields occurrences of any form of the word *słuchać* and a form in the genitive case with one intervening word possible:

```
<lemma=słuchać> <pos=.*:gen:.*> (Slop=1)
```

#	Left	Match	Right
1	i właśnie nie wiem dlaczego nie ale czytałem że jak się jeździ na dużo koncertów właśnie i jak się to przyjmuje i się	słucha dużo muzy	głośno to później się już nie czuje takiego
2	ciocia	słucha Radia	Maryja
3	i chce sobie kupić sobie tego takiego jamnika będzie se stało koło yyyy łóżka mojego będę sobie	słuchać	muzyki kiedy będę chciała nie no mi się to podoba
4	no i	słuchaj do końca	i był sobie ten Leszek co Leszek Leszka w ogóle widywaliśmy mama w takich dziwnych miejscach że wiesz idziesz nagle parkiem i tu Leszek
5	co ty tam nie	słuchaj nikogo	nie słuchaj nikogo
6	nie Zuźka jest odporna jak	słuchaj od	Marcela i Patrycji się nie zaraziła a one non stop chore są w domu osiemnaście stopni mam a ze na wierzchu śpi nogi jak lodek zimne
7	no i	słuchaj poprosiłam studentów	z pierwszego roku no i parę osób mi wysłało no i tak jak mi parę osób wysłało no to wiesz

REST API

The REST API of Spokes PL makes it possible to search and extract the entire contents of the corpus.

- To get the complete list of transcriptions see this link
- Here is how you can get the |list of all utterance turns in this text.

From:
https://212.191.73.246/ - **PELCRA**

Permanent link:
https://212.191.73.246/doku.php?id=spokes_documentation

Last update: **2023/08/18 13:19**