Marek Łaziński, Dorota Górnicka-Urban (UW), Michał Woźniak (IPL PAS)

Tagging aspect pairs in the German-Polish Parallel Corpus

Slavic aspect and (diachronic) corpora
Mainz – March, 29-31, 2021
Founded by NCN and DFG: Beethoven
2016/23/G/HS2/00922

Outline of the presentation

- Definition of an aspect pair
- Aspect pairs in Polish dictionaries
- Tagging aspect in existing Slavic corpora
- Polish-German parallel corpus
- Tagging aspect pairs in the PGPC
  - Automatic tagging
  - Disambiguation by hand
- What we need aspect pair tagging for? Aspect profiles - ratio of suffixal and prefixal opposiitions in Russian and Polish corpora

Aspect pair

Aspect partners must have identical lexixal meaning. The pf partner must denote an event

- Pf ingressiva change the meaning, pf delimitativa do denote events.
- Not every aspect pair is telic, not every telic opposition is an aspect pair (look for - find).

Two verbs create an aspect pair if the pf verb can be replaced by its impf partner in the iterative use or in the past-tense – praesens historicum. (Maslov 1948)

- *Pisze listy* can mean (IT):
  - *Napisał list dwa dni temu, wczoraj, i napisze jutro.*
  - He wrote a letter two days ago, yesterday, he will write tomorrow.
- *Pisze list i wysyła go* can mean (PH):
  - Napisał list, poszedł na pocztę i wysłał go.
  - He wrote a letter, went to the post office and sent it.

Formal criterion

- A prefixal perfective is the aspect partner of a simplex imperfective, if it (the pf) does not derive secondary imperfectives.
- pisać – napisać - *napisywać
- pisać | przepisać – przepisywać

But for triples:

- tworzyć ‚create' – stworzyć – stwarzać
- tworzyć – utworzyć - *utwarzać – only a pair
- tworzyć | przetworzyć ‚transform' – przetwarzać

(| – meaning change)

Pairs in dictionaries: stwarzać – stworzyć, tworzyć – utworzyć, przetwarzać – przetworzyć (no triples).

From triples to pairs. An outline history of the aspect system

- Prefixation:
  - Pisać – napisać ‚write', przepisać – przepisywać ‚copy'
- Secondary imperfective:
  - Pisać - napisać – napisywać
  - Pisać - przepisać – przepisywać
- Loss of secondary imperfectives which repeats the simplex meaning:
  - Pisać – napisać -0
  - Pisać | meaning change przepisać – przepisywać
- Some triples remained:
  - Tworzyć – stworzyć - stwarzać

Aspect pairs in dictionaries

- 17th and 18th centuries: aspect as verb inflexion
- 19th century: aspect as Aktionsart (internal time relation)
- 20th century: all verbs marked for aspect, suffixal pairs described together, no relation or cross-reference in prefixal pairs
- Contemporary dictionaries (after 1990): suffixal and prefixal oppositions described the same way

Vilna dictionary (Zdanowicz et al. 1861)
Warsaw dictionary (Karłowicz et al. 1900-27)
Doroszewski's dictionary (1958-64)

- Every verb in a separate entry, but for suffixal pairs.
- Prefixal pairs without cross reference (with some exceptions).
- Pf aspect partners defined by the ipf
  - *napisać* ‚zapełnić pisaniem' – to fill with writing (Swil)
  - *napisać* ‚skończyć pisanie' – to finish writing (Swar)

Słownik współczesnego języka polskiego (1996), *Inny słownik języka polskiego* (2000)

- SWJP (1996) was the first dictionary describing prefixal and suffixal pairs in single entries with cross references from pfs to ipfs:
  - *gubić – zgubić*, schodzić z drogi – ‚deviate from the way'
  - *strzelać – strzelić,* oddawać strzał – ‚to fire a shot'
  - The ipf achievement verbs in explicatons sound odd lose their meaning of a single event.
- In ISJP (2000), suffixal pairs described in one entry, prefixal pairs in separate entries with cross references from ipfs to the pfs. Opposite to SWJP, the prototype value of aspect is perfective: *zgubić*, *strzelić*.

Verbal Aspect in Russian and Polish dictionaries

- Beginning from Ushakov's dictionary (1936) aspect pairs in Russian dictionaries have been presented according to principles of aspectology treating prefixal and suffixal pairs in the same way.
- Polish dictionaries made this choice only in 1996.

Tagging aspect value in Slavic corpora

- In Slavic corpora every verb form is tagged as ipf or pf without any reference to the aspect pair. Imperfectiva tantum, eg. *być, spacerować* share the same annotation with imperfective aspect partners, eg. *pisać* (opposed to perf. *napisać*).
- Most national corpora of Slavic languages, provide 3 values od aspect: imperf., perf., and biaspectual.
- There is no tag for biaspectual in Bulgarian, Ukrainian, and Polish corpora (biaspectuals are absolute exception among Polish verbs).

Tagging aspect pairs. How?

- The idea of referring verbs to aspect partners in corpora results from their description in dictionaries.
- Relation to aspect partner should be assigned to a specific verb occurrence in the corpus, concerning a lexical sub-meaning, aspectual meaning and context.
- This is only possible when tagging by hand.
- Since we cannot tag the entire corpus by hand. First, every verb occurrance will be tagged alternatively as a partner of different pairs.

Alternative assignment to aspect pairs. *Charakteryzować* ‚characterise / make up'

1. State – ipf tantum:

- Te papugi charakteryzuje piękne ubarwienie.
- 'These parrots are characterised for their beautiful colouring'

2. Accomplishment (telic action) ‚to describe':

- (**S**)Charakteryzował kolegę krótkimi słowami.
- 'He described his friend in short words.'

3. Accomplishment (telic action) ‚make up oneself':

- Aktor (**u**)charakteryzował się przed występem.
- The actor made himself up before the performance.

Every form of the ipf *charakteryzować* must be potentially assigned in the corpus to two different pfs or left as an imperfectivum tantum.

Polish German Parallel Corpus (parasolcorpus.org/Teesthoven/#!/)

- 10 mil. tokens - 2,5 mil. Polish and 2,5 mil. German original texts.
- Time span 1750-2020 (after 1946 2,5 mil.)
- Fiction (55%), non fiction, sub-corpus of law texts (20%)
- Morphological tagging according to Polish National Corpus tagset and SSTS for German.
- Search for word forms, lemmas, grammatical categories (among them general aspect value)
- Verbs tagged as aspect partners

Corpus structure

	POL>GER (5 mil.)	GER>POL (5 mil.)
	Fiction 55%, non-fiction 45%	
1750-1799	0,75 mil. 60% Fiction, 40% non-fiction	0,75 mil. 60% F, 40% nF
1800-1849	1 mil. 60% F, 40% nF	1 mil. 60% F, 40% nF
1849-1899	1 mil. 60% F, 40% nF	1 mil. 60% F, 40% nF
1900-1949	1 mil. 60% F, 40% nF	1 mil. 60% F, 40% nF
1950 -2020 *(copyright problems)*	1,25 mil. 40% F, 60% nF	1,25 mil. 40% F, 40% nF

Aspect tags in the Polish-German Corpus – applied in the search syntax

Aspect value:

- Pf partner, ipf partner, bi-aspectual,
- Verbs not tagged as partners are pf or ipf tantum

Aspect determiner:

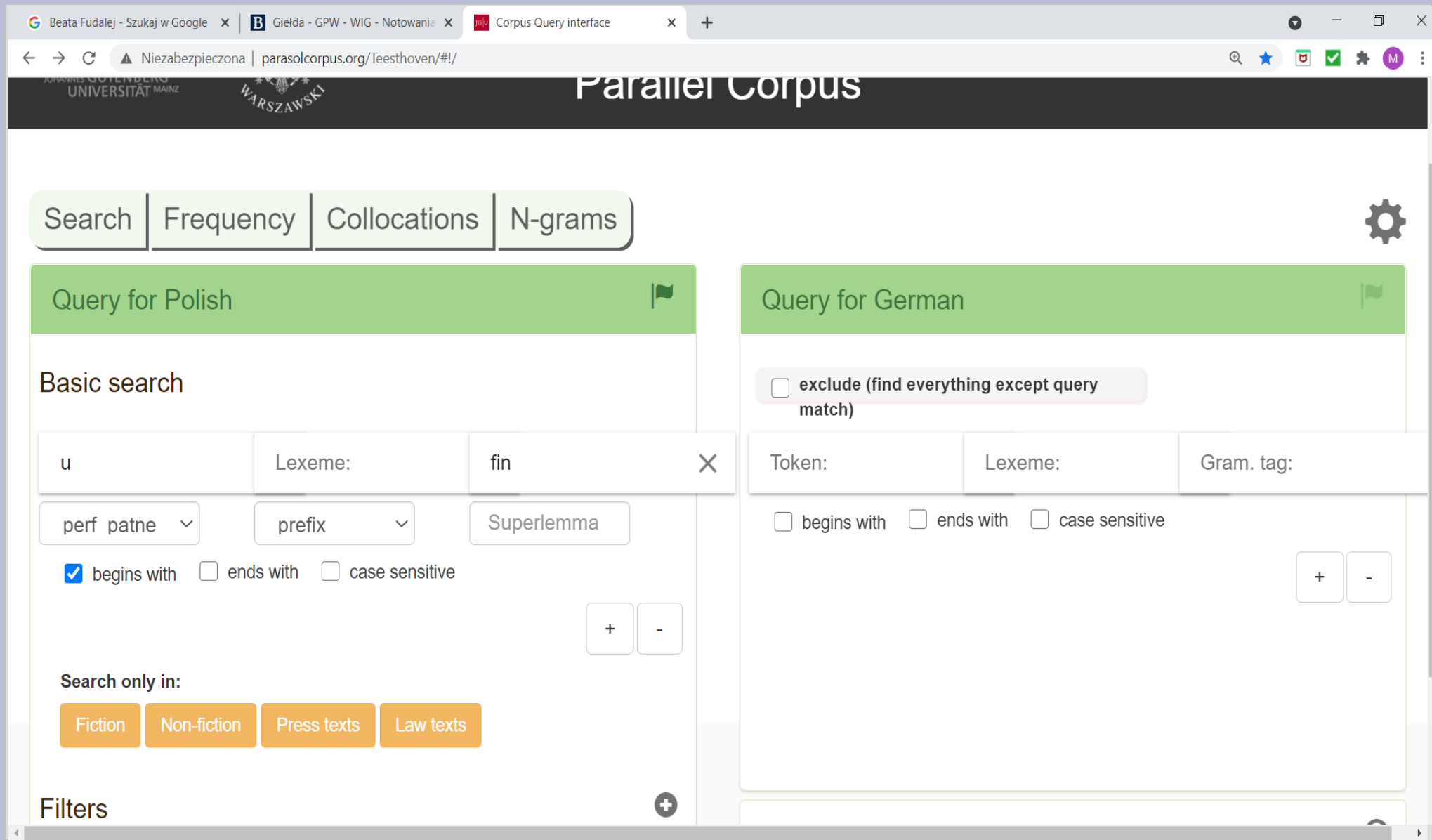
- Not-affixal (in the current interface „simplex”): aspect marked by root, eg. ipf *pisać*, pf *przepisać* (*prze-* is not an aspect prefix, repeated in *przepisywać*)
- Prefix (pf): *na-pisa-ć*
- Suffix: ipf *przepis-ywa-ć*, pf *stuk-nq-ć*
- Suppletive: *brać – wziąć*

Every aspect partner is assigned to a superlemma consisting of ipf and pf, eg. *pisać – napisać.*

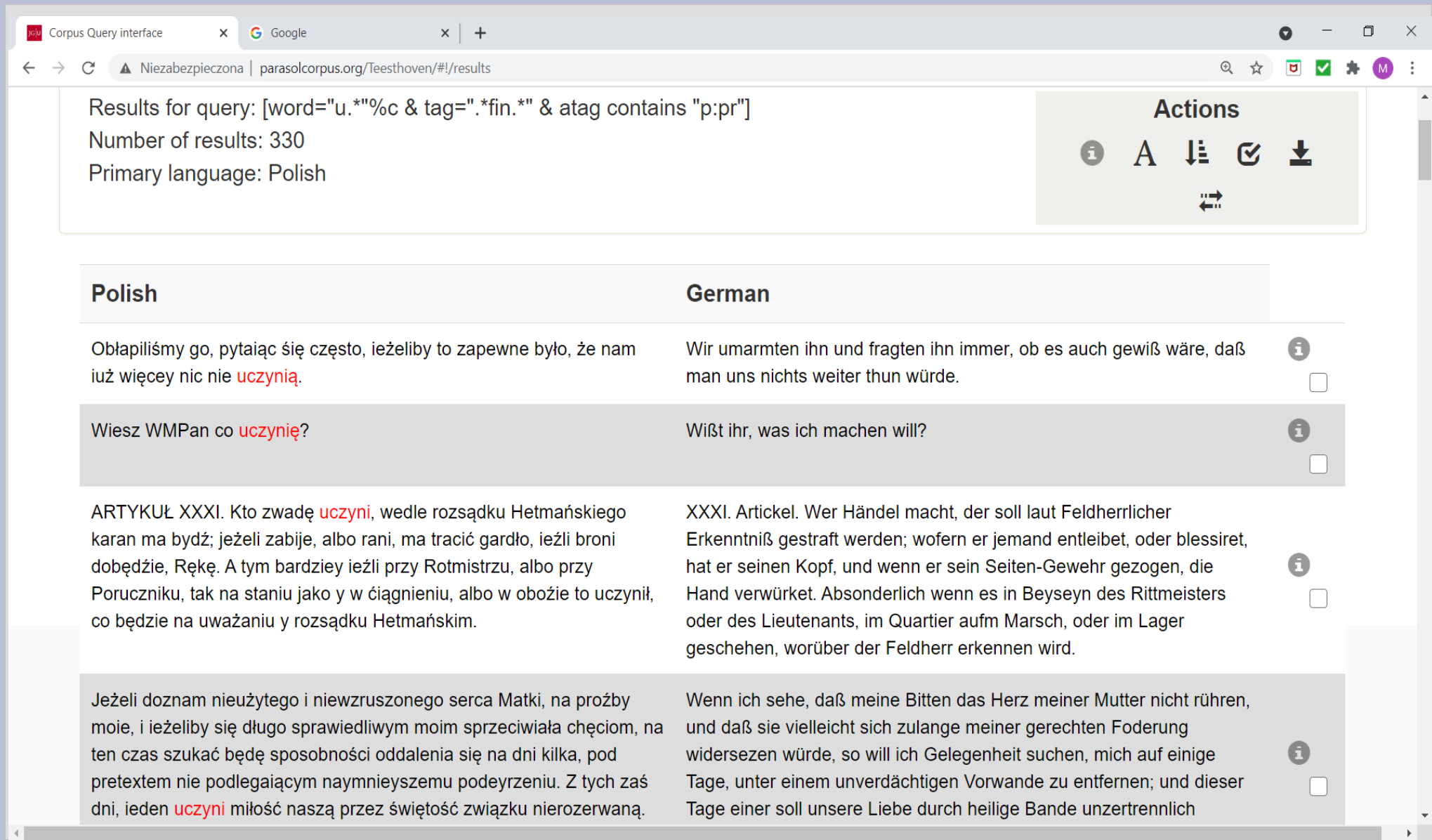
Tagging procedure

- Based on the manually prepared list of Polish verbs (above 10000 entries)
- Annotations assigned automatically basing on lemmas; in case of ambiguity verb is annotated with all possible values.
- Currently in the corpus there are ca. 760 000 verbs of which near 354 000 (47%) are tagged as aspect partners
- The next stage of the planned work is tagging verbs with aspectual meanings nad actional classes.

Search for pf corpus partners beginning with u- in simple future form)



Search results



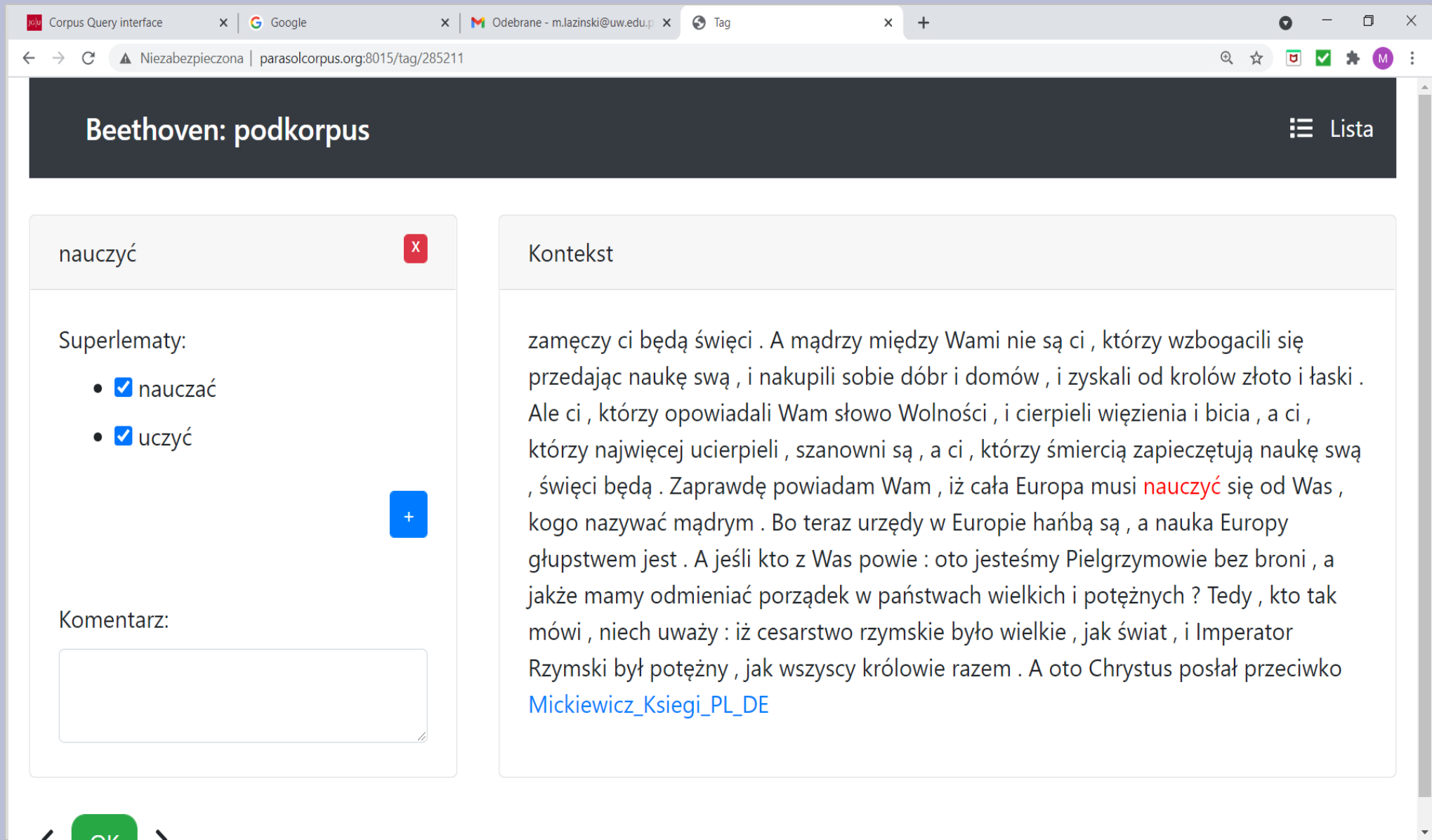
Disambiguation by hand

Automatic choice of the aspect partner will be disambiguated by hand in a sub-corpus containing 45.000 aspect partners. The annotator considers the meaning of the verb in a given context.

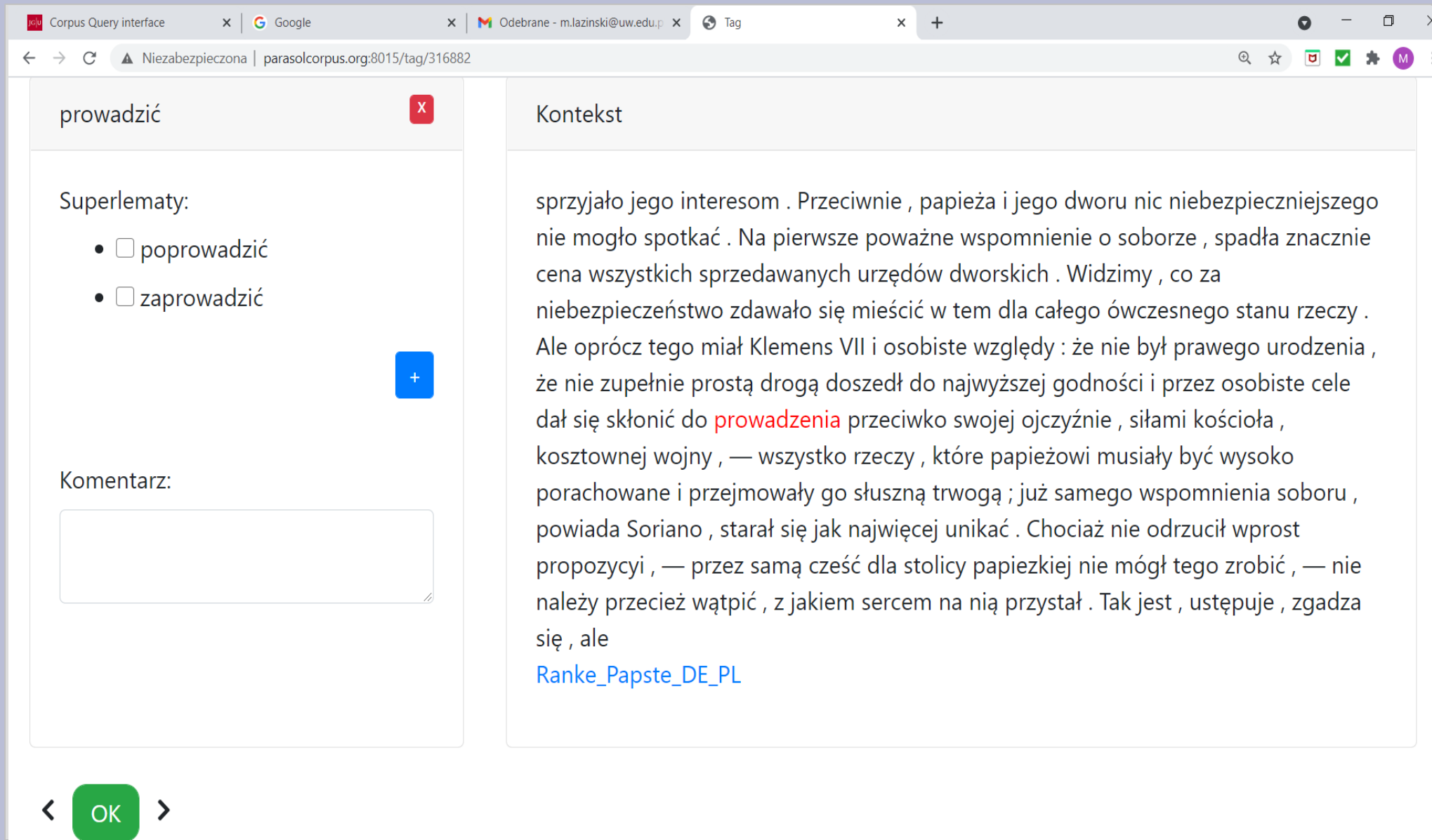
The annotation can be preserved or changed by:

- Removing partners and leaving the verb as ipf or pf tantum.
- Removing one of alternative partners
- Adding a partner
- Changing a partner

Disambiguation interface
Both ipf partners of nauczyć stay and form an aspect triple



Prowadzić wojnę ‚wage a war’ is ipf tantum, potential pf partners: *poprowadzić* and *zaprowadzić* must be removed



Examples

1. […] nie **cierpię** ja jednak wszelkich cicho stąpających męskich nóg. (Nietzsche, Tako rzecze… 1906)

Automatic assignment: cierpieć : ucierpieć. **Decision** – remove ucierpieć (cierpieć ipf tantum)

2. […] to jest ten kompleks zagadnień , który najlepiej chyba **przystaje** do diagnozy Janusza Reitera. (Dialog 2011)

Automatic assignment: przystawać : przystanąć or przystawać : przystać. **Decision** – przystawać : przystać

3. Zeno w swoich naukach o cnocie , **uczy** nas , abyśmy przyrodzone pobudki y skłonności przytłumiali. (Gellert: *Moralne pisma* 1775)

Automatic assignment: uczyć : nauczyć. **Decision** – OK (nauczać has no relation to uczyć)

So far

The sub-corpus contains 52464 aspect partners (tokens):

- 10% from the period 1750-1800, 11% 1750-1800, 12% 1851-1917, 13% 1918-1945,
- 30% from the period 1946-1989, 24% 1990-2021

Up till now 995 partners has been checked:

- In 145 cases the verb has been moved to the class of ipf tantum, eg. *ciepieć* ‚suffer' – *ucierpieć* > remove *ucierpieć*
- In 66 cases the number of potential partners has been resticted, eg. *przystawać* has either the partner *przystać* ‚agree' or *przystanąć* ‚stop'. Never both in the same context.
- In the most cases – 784 the automatical tagging needed no hand correction.

Potential triples with simplex verbs

In older texts, a potential meaning of a simplex verb makes it possible to join it to a suffixal aspect pair where ipf and pf partners contain a prefix specializing new lexical meaning.

- Zwyczay i gust dawny u Turków bywał **lać** armaty wielkie i długie. (Mikosza: Obserwacye 1787)
- Automatic assignment: lać ipf tantum. **Decision –** an aspectual triple lać : odlać : odlewać
- In today's Polish the meaning ‚to cast a cannon' is represented by a verb pair *odlewać – odlać*, the simplex verb *lać* , poor' has not preserved this old meaning. Only the pair *odlewać – odlać*, not *lać - odlać* should be recorded in a contemporary dictionary.

Tagging aspect pairs in a corpus. What for?

- A corpus, tagged for aspect pairs, opens new perspectives for research, especially rich in a parallel corpus with one Slavic and one aspectless language.
- It is possible to establish a ratio of pf and ipf in general and the ratio of all aspect partners.
- Aspect partner value can be combined in queries with other grammatical categories in order to establish the syntactic relation between aspect and these categories.

Ipf : pf aspect partners ratio

In most corpora the ratio ipf:perf is between 2 and 1,5.

- In Polish part of PGPC – 420000 : 280000
- In Polish National Corpus 25,5 mil. : 13,5 mil.
- Among 25,5 mil. Ipfs in PNC, 10 most frequent imperfectiva tantum: *być*, *mieć* etc. make not less than 10,5 mil. words.

The ratio ipf:perf among aspect partners is only 0,8

- 156000:199.000 (pf:ipf = 2,3)
- Query in Pol.-Germ: [atag contains "i(:.*)? "] (or …" p(:.*)? ")

Pf aspect partners outnumber ipf partners in an average text due to the basic narrative function of the pf.

- In order to check this, we need aspect pairs tagged.

Profiles of pf and ipf aspect partners in suffixal and prefixal pairs

Janda and Lyashevksya (2011) proved in a corpus study that prefixal pairs and suffixal aspectual pairs behave the same way on the morphosyntactic level.

- (Aspectual Pairs in the Russian National Corpus, *ScandoSlavica* 57 (2)

We have proved the same for Polish past and non-past verb foms in the Polish-German corpus

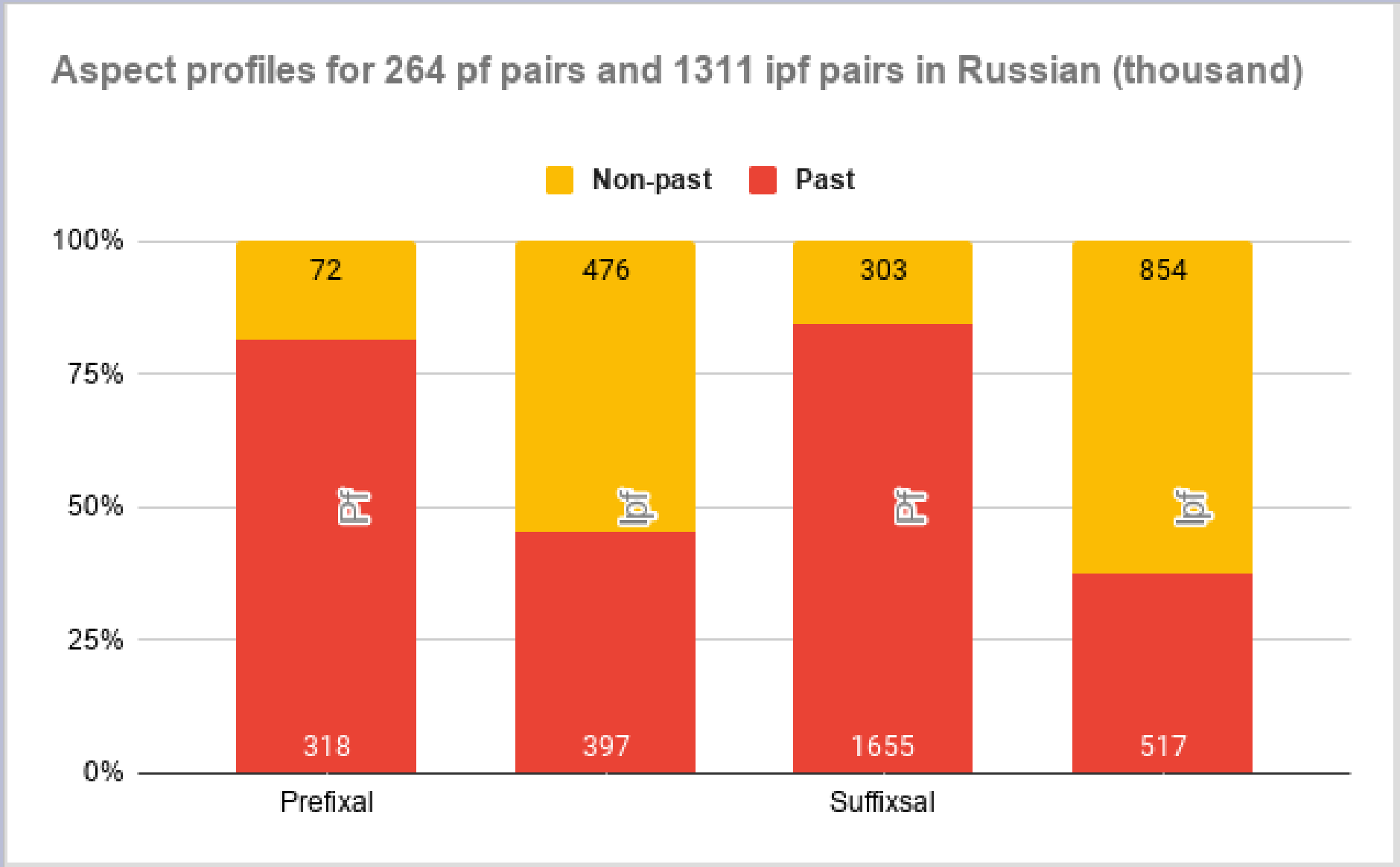
- For example, in order to count prefixal pfs in past tense, aspect value, aspect determiner and morphological tags were combined: [tag=".*praet.*" & atag contains "p:pr"]
- Data in a former version of the corpus with 800 aspect pairs tagged.

A in the Russian corpus, the ratio of aspect partners showed different for different verbal forms but similar for suffixal and prefixal pairs.

Aspect profiles for 264 prefixal and 1311 suffixal pairs in Russian (thousand)

Aspect partners	**Prefixal**			**Suffixal**			**<Sum**
	Pf	Ipf	<Sum	Pf	Ipf	<Sum	
Past	318=82%	397=45%	<715=57%	1655 =85%	517 =38%	<2172= 65%	2887 =63%
Non-past	72=18%	476=55%	<548=43%	303 =15%	854 =64%	<1157 =35%	1705 =37%
Sum	390	873	<^1263	1958	1371	<^3329	<^4592

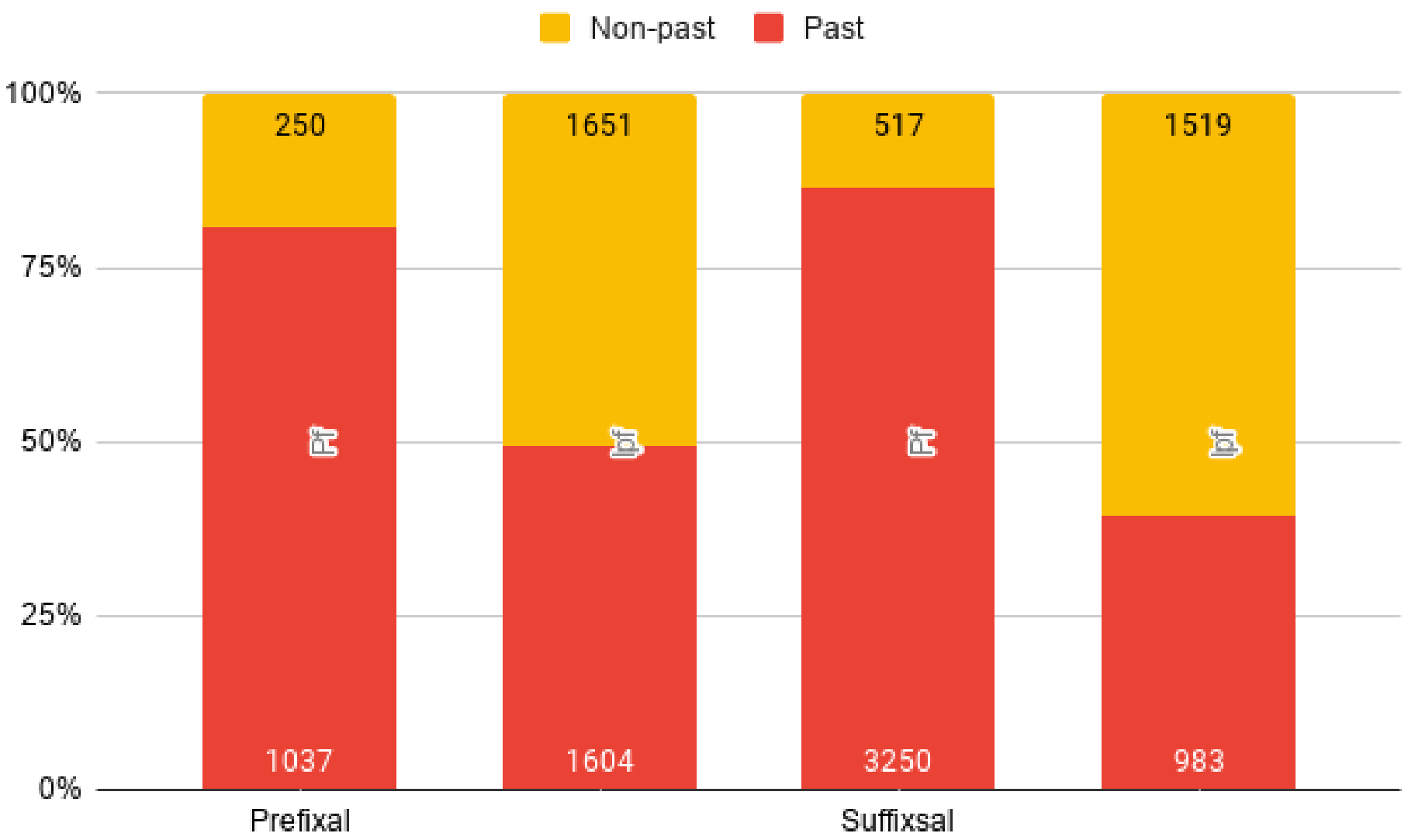
Aspect profiles for 264 prefixal and 1311 suffixal pairs in Russian (thousand)



Aspect profiles for Polish

Aspect partners	**Prefixal**			**Suffixal**			**<Sum**
	Pf	Ipf	<Sum	Pf	Ipf	<Sum	
Past	1037 =81%	1604 =49%	<2641=59%	3250 =86%	983 =39%	<4233 =68%	6874 =64%
Non-past	250=19%	1651 =51%	<1801=41%	517 =14%	1519 =61%	<2036 =32%	3837 =36%
Sum	1287	3255	<^4442	3767	2502	<^6269	<^10711

Aspect profiles for 800 pairs in Polish



Conclusions and future research

- The similarity of aspect profiles counted in a big Russian corpus and in a small Polish one does not end the long-standing discussion about whether the aspectual relation in suffixal and prefixal pairs is the same or not. The search in a corpus tagged for aspect pairs brings an important argument for the identity of two formal kinds of aspect opposition.
- It also shows why and what for a corpus should be tagged up to aspect partners.
- More statistics of verbal aspect in Łaziński: Wykłady o aspekcie: bijp.uw.edu.pl/files/Lazinski2020.pdf