

Grzegorz Murzynowski
natror@o2.pl

Adam Przepiórkowski
adamp@ipipan.waw.pl

Krótką instrukcja anotacji w NKJP

wersja 1.0
26 czerwca 2009 (godz. 03:04)

1. Wstęp

Niniejszy dokument jest krótką instrukcją ręcznej wielopoziomowej anotacji wybranego podkorpusu w projekcie Narodowy Korpus Języka Polskiego.

Staraliśmy się, by ta instrukcja zawierała najważniejsze informacje, lecz by nie była zbyt długa. Dlatego być może brakuje w niej pewnych istotnych informacji. Bardzo prosimy o przesyłanie pytań, wątpliwości i uwag, które mogą się przyczynić do zwiększenia precyzji instrukcji na listę dyskusyjną Anotatorni:

`anot@bach.ipipan.waw.pl`

2. Obsługa programu

2.1. Połączenie z programem

Do anotacji tekstów służy system Anotatornia, z którego korzysta się za pomocą przeglądarki internetowej. W zasadzie każda przeglądarka powinna się do tego nadawać, ale system był intensywnie testowany z przeglądarką Firefox, więc bardzo prosimy o używanie właśnie jej. Można ją pobrać z adresu: <http://www.mozilla-europe.org/pl/products/firefox/>.

UWAGA: w przeglądarce powinna być włączona obsługa JavaScriptu. W wypadku Firefoksa oznacza to, że powinna być zaznaczona opcja „Włącz obsługę języka JavaScript” w menu: Edycja → Preferencje → (zakładka) Treść.

Aby korzystać z programu należy się połączyć z adresem:

`http://chopin.ipipan.waw.pl:8003/`

(Anotatornia produkcyjna¹) lub, do ćwiczeń i wprawek,

`http://chopin.ipipan.waw.pl:8004/`

który różni się od poprzedniego tylko ostatnią cyfrą: 3 — Anotatornia produkcyjna, 4 — Anotatornia ćwiczebno-testowa (dla lepszego odróżnienia wersji ta ostatnia podświetla swoją deskrypcję na czerwono).

Anotatornia pod adresem ...8004 ma służyć użytkownikom Anotatorni do ćwiczeń, zwłaszcza przed rozpoczęciem anotacji na poziomie nie udostępnionym do tej pory (a także do testowania rzeczy nowo zaimplementowanych). Zawiera te same dane

¹ Czyż nie piękne określenie? Takie socjalistyczne, w stylu wczesnego Gomółki: „Anotacji morfoskładniowej ozimej wykunaliiżmy 153 koma 7 kwintala na hektar...”. Ale tak się mówi.

co wersja produkcyjna, ale prawie na pewno transze w wersji ćwiczebnej będą przydzielane w innej kolejności niż w produkcyjnej.

Nowi Anotatorzy otrzymują **Login** i **Hasło** ode mnie (AP). Anotatorki, które brały udział w testach, zachowują swój login i hasło z ostatnich testów (tych pod adresem ...8006). **Login** i **Hasło** są te same w wersji produkcyjnej i ćwiczebnej.

2.2. Transze

Tekst do anotacji jest podzielony na akapity, które zostały pogrupowane w transze po 10 akapitów. Działalność Anotatorki zaczyna się więc od pobrania transzy. W tym celu należy kliknąć: **[pobierz transzę]**².

System pozwoli na czterokrotne powtórzenie tej operacji, a więc możliwe jest pracowanie jednocześnie na pięciu transzach, ale odradzam korzystanie z tej opcji: proponuję pracować na jednej, dokończyć ją, a dopiero potem pobrać następną.

Po przydzieleniu transzy pojawi się jej numer i napis **[pokaż]**. Kliknięcie tego napisu spowoduje wyświetlenie listy akapitów z informacjami o ich statusie – na początku dla każdego akapitu tym statusem będzie **Anotuj**. W wyniku kliknięcia **Anotuj** dany akapit zostaje udostępniony do anotacji.

Po oznakowaniu i zatwierdzeniu akapitu, akapit ten zostaje przesunięty na koniec listy. Natomiast akapity, w których wystąpiły rozbieżności anotacji, umieszczane są na początku listy – sytuacja ta jest sygnalizowana odpowiednim komunikatem. W takim wypadku należy wrócić do danego akapitu, porównać swoją anotację z anotacją drugiej Anotatorki i dokonać odpowiednich zmian lub uprzeć się przy swoim zdaniu. W takiej sytuacji bardzo proszę o wyjaśnianie swojej decyzji w komentarzu.

UWAGA: w wypadku konfliktu nie należy ślepo akceptować decyzji drugiej Anotatorki, lecz należy się dobrze zastanowić nad właściwą anotacją!

Miejsce, w którym wystąpiła rozbieżność, jest podświetlone na żółto, przy czym anotacja partnerki nie jest widoczna.

Akapity powtarzają się w *różnych* transzach, mianowicie każdy w dokładnie dwóch. Żadne dwie Anotatorki nie mają transzy o tym samym numerze. Informacja, kto drugi ma dany akapit i w której transzy, pojawia się na stronie anotacji akapitu, w drugim nawiasie między kropkowanymi liniami (patrz ilustracja 1).

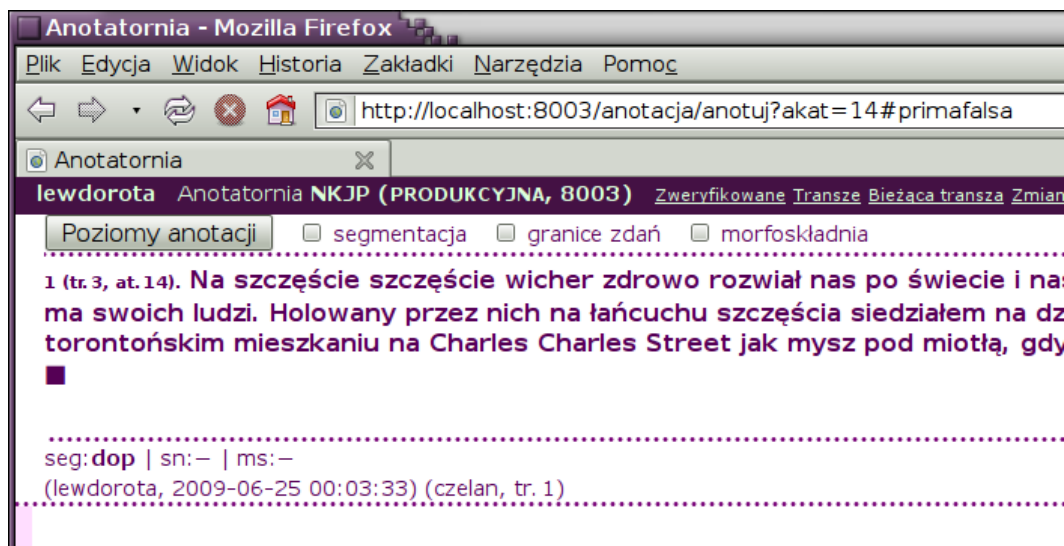
Na stronie z listą transz przy każdej transzy znajduje się informacja w rodzaju:

7 (10 ak., w tym [6, 0, 0] zak. i [1, 0, 0] ocz. – anotuj [3, 6, 6]) **[pokaż]**

»7« to numer transzy. »10« – liczba akapitów w transzy, zwykle równa właśnie 10, ale dla końcowych transz danego etapu wlewania danych może być mniejsza. »[6, 0, 0] zak.« – 6 akapitów zakończonych na poziomie segmentacji i żadnych zakończonych na poziomach granic zdań i morfoskładni. »[1, 0, 0] ocz.« – 1 akapit oczekuje na partnerkę na poziomie segmentacji, na poziomach granic zdań i morfoskładni nic nie czeka. »anotuj [3, 6, 6]« – obecnie masz w tej transzy 3 akapity do ujednoznacznienia segmentacji, 6 do oznaczenia granic zdań i tyleż do anotacji morfoskładniowej. »[przejdź]« jest hiperłączem do tej transzy, po kliknięciu w nie Anotatornia wyświetli listę akapitów tej transzy.

Może też pojawić się informacja o tym, że są akapity do poprawki, w postaci napisu »2 do poprawki« – »2« jest tutaj liczbą akapitów, które należy poprawić (a nie

² O ile pojawia się także napis **[prośba o odebranie transzy]**, należy go ignorować. Operacja ta dotyczy sytuacji, gdy Anotatorka z przyczyn losowych nie jest w stanie dokończyć rozpoczętej transzy.



Ilustr. 1. Fragment strony anotacji. Zwróćcie uwagę na informacje o wyświetlanym akapicie: jest to akapit nr **1**, z transzy **tr. 3**; jego drugą instancję ma **czelan** w transzy **tr. 1**; akapit ten jest **dopuszczony** na poziomie **segmentacji** i niedostępny do anotacji na poziomach granic zdania (**sn:-**) i morfoskładniowym (**ms:-**). **lewdorota** to login aktualnej Anotatorki.

numerem akapitu!). Akapity do poprawki będą wyświetlone (po przejściu do listy akapitów danej transzy) jako pierwsze.

2.3. Poziomy anotacji

Anotacja jest podzielona na poziomy: segmentacyjny, granic zdania, morfosyntaktyczny — te działają w obecnej wersji, oraz, docelowo, sensów słów, słów składniowych, bytów nazwanych i składniowy.

Anotacja na poziomie głębszym jest możliwa po uzgodnieniu anotacji obu Anotatorek na wszystkich poziomach płytszych (zgodnie z grafem na ilustr. 2).

Stan anotacji akapitu jest wyświetlony poniżej treści akapitu, w postaci tekstu w rodzaju:

seg:dop | sn:- | ms:- | wsen:- | synw:- | nen:- | syn:-

gdzie (po lewej stronie dwukropka):

„seg” oznacza poziom **segmentacji**,

„sn” (**s**entences) — poziom granic zdań,

„ms” — poziom **m**orfoskładniowy,

„wsd” (**w**ord **s**ense **d**isambiguation) — poziom sensów słów,

„synw” (**s**yntactic **w**ords) — poziom słów składniowych,

„nen” (**n**amed **e**ntities) — poziom bytów nazwanych,

„syn” (**s**yntactic) — poziom anotacji składniowej,

zaś po prawej stronie dwukropka może pojawić się:

„-” — akapit nie dopuszczony do anotacji na tym poziomie,

→ „dop” — akapit dopuszczony na tym poziomie,

„zatw” — zatwierdzony (ale jeszcze nie uzgodniony z partnerką),

→ „dopo” — do poprawki: próba uzgodnienia z partnerką wykazała rozbieżności; akapit taki należy obejrzeć jeszcze raz i zmienić anotację w miejscach podświetlonych na żółto — tak Anotatornia wskazuje miejsce rozbieżności — lub uprzedzić się przy swoim zdaniu i ponownie zatwierdzić akapit na tym poziomie,
 „popo” — po poprawce, tj. Ty naniosłaś poprawkę i zatwierdziłaś ponownie, ale partnerka jeszcze nie,
 „doos” — po próbie drugiego uzgodnienia anotacji rozbieżności pozostały, akapit przechodzi pod sąd Superanotatora,
 „zwer” — zweryfikowany — anotacja na tym poziomie pomyślnie przeszła uzgodnienie z anotacją partnerki,
 „osa” — osądzany — Superanotator zaczął zajmować się tym akapitem, ale jeszcze nie zatwierdził swoich rozstrzygnięć,
 „oso” — osądzony — akapit został na tym poziomie zaanotowany przez Superanotatora.

Statusy oznaczone strzałką wskazują, że powinnaś coś z tym akapitem zrobić. Zresztą, na liście akapitów danej transzy przy akapitach z takimi statusami pojawia się rozkaznik „Anotuj!” lub „Popraw!”.

Zatwierdzenia anotacji na danym poziomie dokonuje się przez kliknięcie przycisku z oznaczeniem poziomu i checkmarkiem, umieszczonego u góry ekranu, po prawej, poniżej treści akapitu.

W momencie kliknięcia tego przycisku Anotatornia przede wszystkim sprawdzi, czy na danym poziomie zostało oznaczone/wybrane wszystko, co powinno zostać oznaczone/wybrane, i być może wyrazi zdziwienie, np. na poziomie granic zdania tym, że nie wszystkie kropki oznaczono jako kończące zdanie. Jeżeli Anotatorka jest pewna swego, niech ponownie naciśnie przycisk zatwierdzenia zdania. Jeżeli nie jest pewna — niech przejrzy anotację i ewentualnie poprawi, po czym zatwierdzi.

Jeżeli anotacja akapitu na danym poziomie przejdzie wstępny test braku błędów, Anotatornia sprawdza, czy akapit na tym poziomie został już zaanotowany przez Twoją partnerkę. Jeśli nie, akapit otrzymuje status „zatwierdzony” i czeka, aż zaanotuje go partnerka, do tego czasu Ty możesz zmieniać jego anotację. Jeżeli akapit został już zaanotowany przez partnerkę, Anotatornia sprawdza zgodność obydwu anotacji na danym poziomie. W razie stwierdzenia niezgodności kieruje akapit do poprawki, a w przypadku zgodności anotacji — uznaje go za zweryfikowany na tym poziomie, i udostępnia go do anotacji na kolejnym poziomie (o ile jest kolejny poziom).

2.3.1. Oglądanie poziomów anotacji

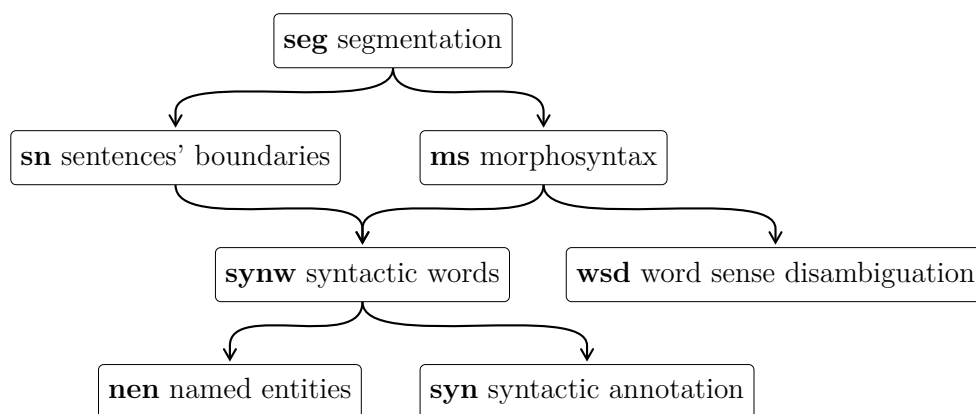
Anotatornia, jeśli nie zaznaczono inaczej, wyświetla ten poziom anotacji (te poziomy), które w danym akapicie trzeba aktualnie anotować.

Uwaga. Gdy akapit czeka na danym poziomie na anotację partnerki (pierwszą lub poprawkową), poziom ten domyślnie nie jest wyświetlany. Możesz jednak go wyświetlić, klikając odpowiednie pola wyboru na górze strony, tuż poniżej fioletowego paska, i zatwierdzając przyciskiem »Poziomy anotacji«.

(Wówczas możesz też zmienić swoją anotację na tym poziomie.)

Taki specjalny wybór poziomów anotacji obowiązuje do następnej zmiany i wylogowanie się nie zmienia go.

Dla akapitu zakończonego Anotatornia domyślnie wyświetla poziom(y) ustalone do wyświetlania w danej wersji programu (2009/05/19 są to segmentacja, granice zdań i morfoskładnia), ich wybór można zmienić jak wyżej.



Ilustr. 2. Porządek poziomów anotacji. Podajemy angielskie deskrypcje poziomów dla lepszego wyjaśnienia ich oznaczeń skrótowych. Strzałki pokazują zależność, a zatem także porządek dopuszczania. Zwróćcie uwagę na to, że sn można anotować niezależnie od ms, oraz synw (docelowo, obecnie jeszcze nie działa) niezależnie od wsd (który ma zadziałać na dniach). wsd jest także niezależny od sn.

2.3.2. Poziom segmentacyjny

Pierwszym poziomem anotacji jest poziom segmentacyjny, na którym Anotatorka ma za zadanie wybrać jeden z wariantów podziału tekstu na segmenty, np. spomiędzy »gdzieś adv« i »gdzie qub, ś być aglt:sg:sec:imperf:nwok«. Spodziewamy się, że potrzeba rozstrzygania niejednoznaczności segmentacyjnej będzie zachodziła rzadko (w jednej z wersji testowych był jeden taki akapit na ok. 300 wszystkich).

Warianty segmentacyjne danego wyboru są podświetlone różnymi odcieniami seledynu. Aby wybrać jeden z wariantów, należy kliknąć radioguzik „wybierz tę segm.” któregośkolwiek jego tokenu, bądź „odrzuć tę segm.” tokenu wariantu, który chcesz odrzucić. Kliknięcie spowoduje dokonanie i zatwierdzenie wyboru, co powinno natychmiast uwidocznic się jako oznaczenie odrzuconego wariantu tłem fioletowym. Podjętą decyzję można zmienić: po kliknięciu w [zmień] znowu pojawi się zestaw radioguzików.

W momencie zatwierdzenia anotacji na tym poziomie Anotatornia sprawdza, czy ujednolicono wszystkie niejednoznaczności segmentacyjne (może się zdarzyć, że w akapicie występuje więcej niż jedna) i, jeżeli nie, nie pozwoli zatwierdzić.

W typowym przypadku, gdy akapit jest jednoznaczny segmentacyjnie, zostaje on na poziomie segmentacji oznaczony jako zweryfikowany automatycznie (w ramach wlewania danych do bazy) i powyższe omija Anotatorkę.

2.3.3. Dodawanie wariantu segmentacyjnego

Anotator sam z siebie nie ma możliwości utworzenia nowego wariantu segmentacyjnego, może jednak zgłosić takie żądanie (prośbę), którą Anotatornia przekaże osobom uprawnionym do jej rozpatrzenia (Superanotatorowi lub Audytorowi):

U dołu strony anotacji akapitu znajduje się link »Zgłoś prośbę o nową segmentację«. Jego kliknięcie spowoduje wyświetlenie formularzyka, w którym należy wpisać co jest (»orth-y, które chcesz podzielić/połączyć...«), co ma być (»proponowany podział/połączenie...«), i ewentualnie dlaczego (»opis«).

UWAGA: formularzyk jest kasztoczuły (case sensitive), na co uczulamy, bo zdarzyło

się w testach, że Anotatorka wpisywała „myśmy” i „my śmy”, gdy w akapicie było „Myśmy” (wówczas Anotatornia powiedziała, że nie widzi takiego tokenu w akapicie).

Osoba uprawniona widzi zgłoszoną prośbę, gdy tylko się zaloguje do Anotatorni, i może ją spełnić lub odrzucić. Spełnienie prośby skutkuje utworzeniem takiego wariantu segmentacyjnego dla *wszystkich* tokenów o podanym orth-u (lub układów tokenów w przypadku łączenia).

2.3.4. Poziom granic zdania

Na tym poziomie zadaniem Anotatorki jest oznaczenie pewnych tokenów jako kończących zdania. W typowym przypadku tokenem kończącym zdanie będzie zapewne znak interpunkcyjny: kropka, pytalnik lub wykrzyknik.

Anotatornia umożliwia oznaczanie granic zdań w dwóch miejscach: w treści akapitu powtórzonej w postaci ciągu hiperłącz, tuż poniżej górnego, nieruchomego panelu (segmenty w ramkach barwy fuksji, po najechaniu myszą tło robi się seledynowe), lub w liście tokenów podanej w postaci długiej kolumny.

Jeżeli Anotatorka woli oznaczać granice zdań w powtórzonej treści akapitu, wystarczy, że kliknie w hiperłącze żadanego tokenu, a zostanie on uznany przez Anotatornię za koniec zdania, co objawi się dodaniem na jego końcu pełnego kwadracika. Kliknięcie tokenu uznanego za koniec zdania, tj. zaopatrzonego w kwadracik na końcu, spowoduje uznanie go za nie kończący zdania, co objawi się usunięciem kwadracika.

Oznaczenie tokenu jako kończącego zdanie jest widoczne także na długiej liście: u dołu pola takiego tokenu pojawia się linia z kwadracików i oznaczenie „(k.z.)”.

Anotatorka może wedle kaprysu bądź uznania oznaczać końce zdania także na długiej liście tokenów, klikając link [tnij za mną] lub [nie tnij za mną], który zmienia końcozdaniowy status tokenu.

Ostatni token każdego akapitu został automatycznie oznaczony jako kończący zdanie, co w szczególnych przypadkach może wymagać zmiany.

Jeżeli granice zdań w akapicie zostały zweryfikowane, to pojawiają się w postaci kwadratów w treści zdania wyświetlanej w górnym, nieruchomym panelu (a ich zmiana przez Anotatora nie będzie możliwa).

2.4. Poziom morfoskładniowy

Dane wejściowe Anotatorni są z założenia niezdezambiguowane i zadaniem Anotatorki na tym poziomie anotacji jest właśnie dezambiguacja morfoskładniowa wszystkich tokenów, z wyjątkiem tokenów interpunkcyjnych, które z założenia mają tylko jedną interpretację morfoskładniową, „interp”.

W polu tokenu, który ma zostać zdezambiguowany, znajdują się dwa linki: [dezambiguuj!] i [dodaj]. Kliknięcie pierwszego z nich spowoduje dodanie do listy proponowanych interpretacji morfosyntaktycznych radioguziki, które umożliwią wybór jednej z interpretacji. Link [dodaj] wyświetla zestaw pól dla dodania interpretacji morfoskładniowej: pole lewe dla lematu (formy hasłowej), pole prawe dla tagu.

Pole tagu wyposażono w mechanizm autouzupełniania, który pozwala zarówno wpisywać tag „z głowy”, jak i tworzyć go przez wybór kolejnych jego członów z listy podpowiedzi: jeżeli coś wpiszesz (i zaczekasz chwilę), poniżej pola tagu rozwinie się lista podpowiedzi zgodnych z tagsetem, po której możesz poruszać się kursorami, a wybierasz tabulatorem lub klawiszem Enter. Jeżeli nic nie jest wpisane i chcesz listę dostępnych klas gramatycznych, wystarczy, że naciśniesz kursor w dół (↓). Kursor w dół rozwija listę podpowiedzi także w trakcie tworzenia tagu, kiedy wybrałeś kolejny człon tagu.

Przy tokenie zdezambiguowanym zamiast [dezambiguuj!] wyświetla się link [wybierz], na wypadek, gdybyś chciała zmienić swoją decyzję — co możesz uczynić aż do momentu weryfikacji akapitu.

Tagi podane do wyboru są z założenia zgodne z Tagsetem NKJP. Tagi tworzone przez Anotatorki są sprawdzane pod kątem zgodności z tymże Tagsetem. Zgodnie z zaleceniem AP „poniższe ograniczenia i zależności są mniej twarde, ale idealnie by było, gdyby próba stworzenia tagów niezgodnych z nimi była lekko szykanowana:”

ger → number = sg

num → number = pl

[trzecie ograniczenie?]

„Lekkie szykanowanie” polega na tym, że na liście autouzupełnienia pojawia się tylko wartość preferowana, jednak nie zostanie zgłoszony błąd i tag zostanie zatwierdzony, jeśli Anotatorka ręcznie wpisze wartość „szykanowaną” (czyli »pl« dla ger lub »sg« dla num). W tym ostatnim przypadku autouzupełnianie podpowiada dalsze części tagu z wartością „szykanowaną”.

Pozostałe ograniczenia zbiorów wartości poszczególnych atrybutów, określone jako „twarde”, są przestrzegane rygorystycznie, tj. np. tag »pant:imperf« (który można by wpisać tylko ręcznie) nie zostanie zatwierdzony, ale wygeneruje komunikat o konieczności poprawki.

Twórca tagów sprawdza także, jeżeli proponowaną klasą gramatyczną jest »siebie«, czy proponowanym lematem jest »siebie« (małymi literami) i odmawia zatwierdzenia tagu, gdy nie jest (w szczególności dla »Siebie« z wielkiej litery).

3. Zgłaszanie błędów

Jak większość produktów działalności ludzkiej, także Anotatornia zawiera pewne niezamierzone niedoskonałości, zwane błędami. Jeżeli użytkownik napotka taki błąd, zwłaszcza gdy uniemożliwiający pracę, powinien go zgłosić na trackerze

<http://chopin.ipipan.waw.pl:8008/Anotatornia>

z zaznaczeniem kto (login), o której godzinie i który akapit(transza, akapit_transzy).

Dane te bardzo się przydają zwłaszcza przy naszych ulubionych błędach „We are sorry but something went wrong”. »akapit(transza, akapit_transzy)« jest to trójka numerów, ostatnie dwa w nawiasie, wyświetlana tuż przed treścią akapitu (patrz zrzut ekranu). Przydatny jest zwłaszcza ostatni w nawiasie, jest to bowiem id instancji akapitu anotowanej/oglądanej przez tego użytkownika.

Wszystko to daje się wygrzebać z logu, więc niepodanie tych danych nie jest krytyczne, ale wydatnie by przyspieszyło rozdeptywanie pluskwy :-).

4. Nieoczywistości — z trackera

PYTANIE: Czy wylogowanie się w trakcie anotacji, tzn. bez zatwierdzenia akapitu, spowoduje utratę wprowadzonych zmian? — pytają Anotatorki co najmniej w dwóch bilecikach ślednika.

ODPOWIEDŹ: Nie, nie spowoduje. Wszystkie zmiany są na bieżąco zapisywane w bazie danych i Anotatornia przechowa je między logowaniami. Co prawda nie zapamięta, który akapit był ostatnio „na tapecie”.

PYTANIE: Po wciśnięciu »wybierz«, zniknął całkowicie cały wiersz z lematem i charakterystyką morfosynt. oraz klawiszki »wybierz« i »dodaj«. Zostało tylko zapisane kursywą (czyli segment) »Jeśli«. Tak powinno być?

ODPOWIEDŹ: Jest to przykład błędu, i bardzo dobrze, że został zgłoszony. Proszę i nalegam, żeby zgłaszać *wszelkie* nieprawidłowości i nieoczekiwane zachowania Anotatorni. Natomiast w pewnych przypadkach można poradzić sobie samej, np. klikając »odśwież stronę« (w Firefoxie także Ctrl-R). Zresztą widzę po logu i stanie bazy, że w tym wypadku Anotatorka poradziła sobie znakomicie, zanim poprawiłem ten błąd (bo zaanotowała akapit).

— Ponownie zgłaszam kwestię...

ODPOWIEDŹ: Proszę raczej spojrzeć na status tego bileciku, w którym sprawa została zgłoszona po raz pierwszy, i ewentualnie dopisywać komentarze do owego pierwszego bileciku w danej sprawie — który powinien pozostać jedyny.

Zresztą, zgodnie z „dobrą praktyką morską”, zanim samemu się zgłosi jakiś błąd, wypada przejrzeć błędy już zgłoszone, czy taki już się nie pojawił, i ewentualnie dopisać swoje „trzy grosze” do już istniejącego bileciku.