https://doi.org/10.23939/sisn2019.01.101

УДК 811.162.1=162.2'374.822:004.65

USING TRANSITIVITY INFORMATION FOR MORPHOLOGICAL AND SYNTACTIC DISAMBIGUATION OF PRONOUNS IN UKRAINIAN

Natalia Kotsyba[1,2], Bohdan Moskalevskyi[2]

[1]Samsung Research Poland, pl. Europejski 1, Warsaw, Poland

[2]Institute for Ukrainian, NGO, 27, Holovka Str., Kyiv, Ukraine

[1]E-mail: gnatko@gmail.com, ORCID: 0000-0002-1230-8788;

[2]E-mail: msklvsk@icloud.com, ORCID: 0000-0002-1803-9806

The paper presents a short introduction to several electronic resources for Ukrainian language, namely, two treebanks: the Gold standard (ab. 130 thousand tokens), manually annotated in the Universal Dependencies flavour (https://universaldependencies.org/), which comprises the training data for a machine-trained syntactic parser, and a big (near 3 billion tokens), automatically annotated General Treebank (also known as Zvidusil), as well as a valency dictionary, developed by the Institute for Ukrainian, NGO (Kyiv) in 2015-2019 (https://mova.institute/). We also describe an experimental usage of the valency dictionary information to boost the performance of the syntactic parser. As a proof of concept, we discuss the case of syntactic and morphological ambiguity of frequently used Ukrainian pronouns його, її, їх 'his, her, their' and ways of improving the syntactic parser's performance using the supervised machine learning techniques with a theoretical linguistic support. Apart from the multiple morphological ambiguity (24+ possible tags for each of these forms), one of the challenges connected with the presented linguistic phenomenon, is that its correct disambiguation involves anaphora resolution and semantic roles identification. On the one hand, this makes the disambiguation process much more complicated, given the followed annotation design, on the other hand, by resolving a seemingly low-level (morphological) problem we gain a bonus in the form of significant textual analysis hints which can be later used in various NLP applications for Ukrainian. The present article is a practical follow-up of its more theoretical predecessor (Kotsyba, Moskalevskyi 2018 [11]), where the linguistic underpinnings of the syntactic and morphological interpretation of the pronouns його, її, їх in comparison with other Slavic languages are presented in greater detail.

Key words: Ukrainian language, Treebank, syntactic parsing, semantic roles, valency dictionary, anaphora resolution, morphological disambiguation, supervised machine learning.

Наведено короткий опис декількох електронних ресурсів української мови, а саме два синтаксичні корпуси: Золотий стандарт (біля 130 тис. слів), анотований вручну деревами залежностей Universal Dependencies (https://universaldependencies.org/), що становить тренувальні дані для синтаксичного парсера, та великий (майже 3 мільярди слів) автоматично анотований Загальний синтаксичний корпус (Звідусіль), а також валентний словник українських дієслів. Ці мовні ресурси розробляються в Інституті Української, ГО від 2015 року та є доступні для некомерційного вживання під адресою установи https://mova.institute/. Також описано експериметальне використання валентного словника для покращення якості роботи синтаксичного парсера з використанням машинного навчання та ґрунтовної теоретико-лінгвістичної бази. Прикладом були конструкції особово-присвійних займенників "його", "її", "їх", кожен з яких має понад 24 можливі морфологічні таги, у сполученні з ґерундієвими іменниковими формами, що також можуть мати різні граматичні інтерпретації (із ключовими семантичними ролями або без них). Вибір правильної інтерпретації у багатьох випадках вимагає ідентифікації семантичної ролі іменника, що його заступає у тексті займенник, і/або розв'язання кореференції (анафори). З одного боку, це ускладнює процес

уоднозначнення; з іншого боку, ми отримуємо бонус для якісного автоматичного аналізу тексту, необхідного для багатьох застосувань в обробці природних мов (NLP). Проаналізовано типові помилки автоматичного парсингу для досліджуваної конструкції та подано практичні рекомендації до створення тренінгових даних для кращого навчання парсера у майбутньому. Стаття є практичним продовженням лінгвістичного дослідження (Kotsyba, Moskalevskyi 2018 [11]), де подано теоретичне обгрунтування рішення проблеми інтерпретації займенників та ґерундієвих іменників для української мови на тлі інших словянських мов.

Ключові слова: українська мова, снтаксичний корпус, дерево залежностей, валентний словник, семантичні ролі, анафора, морфологічне уоднозначення, машинне навчання.

Introduction

High accuracy syntactic and morphological parsing still remains one of the biggest challenges of the natural language processing, especially for the morphologically rich languages like Ukrainian. In the present paper we are going to describe some of the disambiguation problems we have encountered while training a syntactic and morphological parser for Ukrainian and possible ways to cope with them. The paper is structured as follows: Sections 2 and 3 give a short overview of the language resources used (composition of treebanks, the syntactic parser's performance, peculiarities of annotation, and the valency dictionary in development), Section 4 describes an experiment conducted to verify how making the parser partially valency aware reflects its performance for the investigated structure, Section 5 is an overview of possible disambiguation solutions, and finally Section 6 presents conclusions and possible further work.

Treebanks and the parser used

The presented below resources are being currently developed by Institute for Ukrainian, NGO as a grass root initiative in partial cooperation with the faculty of philology of Kyiv Mohyla Academy, see also [10]. They are made publicly available for non-commercial use.

The present[1] **IU Gold** standard Treebank ("IU" in its title stands for "Institute for Ukrainian") contains fragments of genre balanced texts from the XX–XXIst centuries, amounting to over 130K **manually** annotated tokens with morphological and syntactic features. The morphological annotation quality was assured by a 2+1 system, where two independent annotators worked on the same texts and the third annotator resolved any discrepancies in their annotation. The syntactic part is being developed within the Universal Dependencies (hence, UD) project[2] since November 2015, which makes it conceptually aligned with other 70+ languages and quality attested. This layer of annotation is done by only one human annotator but later it goes through checks by other annotators and the whole Treebank is subjected to more than two hundred of manually designed and programmed consistency tests.

IU General Treebank (Zvidusil)[3] is parsed **automatically** based on the training data of the Gold standard. It contains 2 848 203 658 tokens, mainly harvested from the Internet or granted by friendly publishing houses. The essential parts of it are: fiction (also translated), newspapers, fora, blogs, manuals, documents. Both Gold and General Treebanks are the first syntactic resources for the Ukrainian language of this size and quality[4].

[1] Starting from of May 2018, both treebanks are searchable through one of the alternative engines: https://mova.institute/kontext or https://mova.institute/bonito. IU Gold Treebank can be downloaded from https://mova.institute or https://github.com/UniversalDependencies/UD_Ukrainian-IU/tree/dev.

[2] https://universaldependencies.org

[3] The original name is **Zvidusil**, from Ukrainian *звідусіль* meaning 'from everywhere'.

[4] Another existing project dedicated to Ukrainian syntax that deserves attention is developed at the Institute of Philology of Kyiv National University (http://www.mova.info/Page2.aspx?l1=14) but as of March 2019 it looks like rather a small, experimental resource as compared to ours: http://www.mova.info/syntaxis_search.aspx

The morphosyntactic and syntactic parser was trained on the Gold standard Treebank and its performance on 15K test set as of May 2018 is summarised in Table 1 below:

Table 1

Statistics on the performance of the UDPipe[5] morphosyntactic parser for Ukrainian

metric	on plain text (%)	pretokenised (%)
universal[6] parts of speech	97.25	97.45
UD morphological features	91.48	91.61
UD whole tags	90.87	91.00
lemmas	98.2	98.43
UAS: unlabelled attachment score (head without relation)	79.27	82.1 (also manually premorphotagged)
LAS: labelled attachment score (head and relation)	75.52	79.89 (also manually premorphotagged)

Even though the parser performs quite well as compared with corpora for other languages in UD, with roughly every 10th morphological and 4th syntactic tag being wrong, its everyday working use for linguistic analysis is still not possible. This is the reason why we are looking for ways to enhance its parsing performance. The analysis of typical parsing mistakes of the IU syntactic parser reveals that it copes well with very frequent phenomena but requires better training with respect to rare ones. Pronouns make one of the problems which deserves special attention due to a high frequency of their use. Some of the contexts they appear in, however, are not so frequent and may cause difficulties for automatic parsing.

Quality of annotation largely depends on the annotation scheme design and the chosen level of its granularity, and to some extent can be manipulated by adjusting both the parameters. However, in the case of the syntactic parser trained on the Gold Treebank, the area for manoeuvres is limited by the accepted international standards of annotation, namely, the scheme used by the Universal Dependencies initiative. The UD project, whose aim is a consistent cross-linguistic syntactic annotation of many languages, started in 2013 and by March 2019 the quantity of languages has grown to 76, with a dozen more upcoming. The Ukrainian branch has been developed there since 2015, and by July 2018 five stable releases had been produced. The annotation scheme used in UD has its roots in the Stanford dependencies for English [3], Google universal part-of-speech tags [15], and the Interset interlingua for morphosyntactic tagsets [18], but it is still constantly evolving to reach better consistency throughout the involved languages. Apart from the necessary common core of the universal parts of speech, features, and relations, each language may have language-specific features as well, so its peculiarity is not compromised for the sake of the general “good”.

The Ukrainian version of the UD tagset has its roots in the MULTEXT-East v.4 (MONDILEX, or MTE for short) morphosyntactic tagset, for more details see [5, 7]. In the light of the present research this relation is important, as the discrepancies in the pronouns conceptualisation in both projects were reflected later on some aspects of their (pronouns) annotation. In particular, MTE was focused rather on the morphological and partly also etymological aspects of pronouns while leaving the syntactic function to be dealt with at the level of syntax.

[5] State-of-the art parsers, e.g. Stanford [4] reach far better performance (up to 87.5 % LAS), but do not support plugging in a morphological dictionary, which is needed for our experiments. We therefore use UDPipe.

[6] In this context “universal” refers to those used in the Universal Dependencies (UD) project.

Table 2

Possible combinations of tags for Ukrainian pronouns in MULTEXT-East v.4 (MONDILEX)[7]

POS	Type	Ref_Type	Person	Gender	Animate	Number	Case	Synt_Type	Example
P	p		12		y	sp	ngdail	n	я, мене, ти, тобі, ми, нами, ви, вас
P	p		3	mfn		s	ngdail	n	він, вона, воно, **його**, **її**, йому, ним, нею, ньому, ній
P	p		3			p	ngdail	n	вони, **їх**, їм, ними, них
P	di**s**qrzgx			m		s	ngdil	a	такий, цього
P	di**s**qrzgx			fn		s	ngdail	a	така, цього, цій, оте
P	di**s**qrzgx					p	ngdil	a	такі, таких, цим
P	di**s**qrzgx			m	yn	s	a	a	того, такого, тих, таких, той
P	di**s**qrzgx				yn	p	a	a	тих, ті

UD has a unified treatment of pronouns with respect to their syntactic environment. Therefore, the forms *його, її, їх* 'his, her, their' receive 24 additional tags each as possessive pronouns to support the noun agreement information in UD, in addition to the original, much fewer, personal pronouns tags in MTE.

Example of pronoun tags in the UD Gold Treebank for Ukrainian 2 out of 29 available tags for the form *його*.

його він PRON Case=Acc|Gender=Masc|Number=Sing|Person=3|PronType=Prs

його його DET Case=Nom|Gender=Masc|Number=Sing|Person=3|Poss=Yes|PronType=Prs|Uninflect=Yes

This multiplicity of tags is a cause of disambiguation trouble in general, although most often it is the basic possessive (DET)::personal (PRON) distinction which is of the most practical significance.

Introduction of the possessive interpretation of *його, її, їх* pronouns demands automatic disambiguation in the cases with which even human annotators themselves may have troubles. Let us discuss some examples first.

In the easiest situation, when these forms precede nouns, they are possessives, while when they precede verbs, they are personal pronouns replacing some of the verbal arguments (most often Agent of Patient). However, there is a specific type of situation when nominal and verbal qualities are mixed, namely, the gerundial form (in academic grammars and dictionaries of Ukrainian these are nouns of the deverbal origin ending with *-ння/-ття*)[8]. Example of the possessive use of the pronoun: [9]

Природно, що соціологія найчастіше повинна розглядати ***індивіда*** *і* ***його*** *положення і значення в різних соціальних зв'язках.*' Naturally, sociology most often has to consider the **individual** and **his/her** position and significance in various social relationships.'

[7] http://nl.ijs.si/ME/V4/msd/html/msd.P-uk.html

[8] Following the way it is done in [11], we will be using the term "gerund-like form" (GLF, for short) to talk about such nouns in a generalised way, before disambiguating them into nouns or gerunds proper.

[9] Demo of the on-the-fly parsing is available at https://mova.institute/аналізатор.

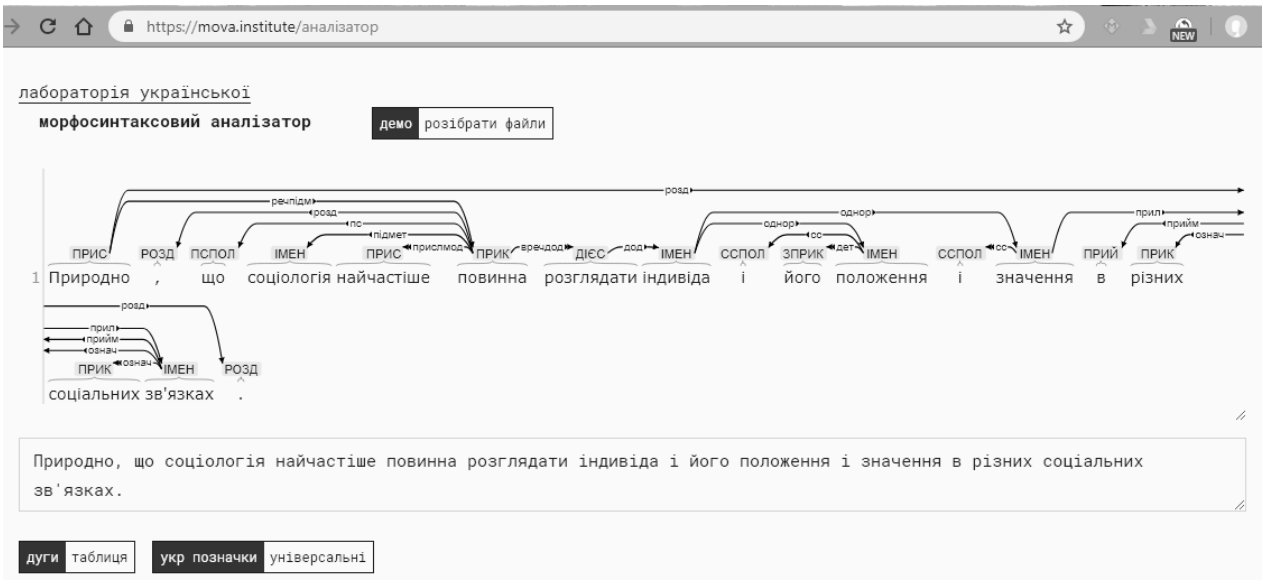


Fig. 1. Parsing results for the possessive pronoun, tree mode

The same sentence parse can be shown in the table format.[10]

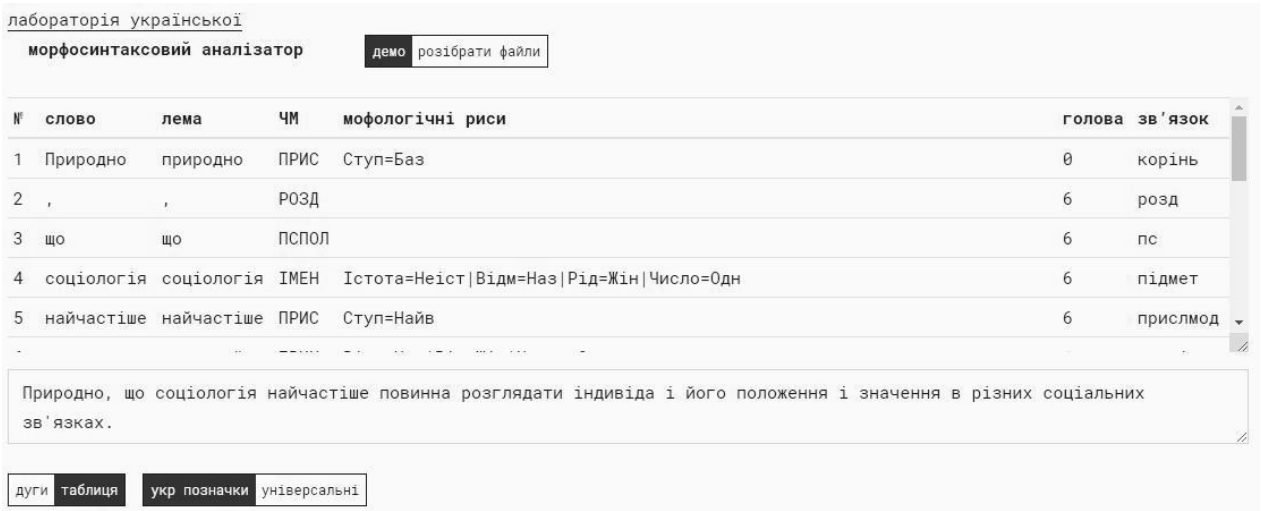


Fig. 2. Parsing results for the possessive pronoun, table mode

Example of the quasi-possessive use of the pronoun, where *його* 'his' refers to the proper name subject "Роберт Мюллер", the Agent of the gerundial action.

Інформацію було передано адвокату ***Роберту Мюллеру*** *в рамках* ***його*** *розслідування*. 'The information was forwarded to the lawyer **Robert Muller** as part of **his** investigation.'

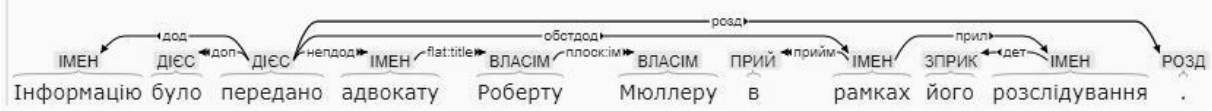


Fig. 3. Parsing results for the Agent pronoun role, tree mode

Such uses are treated as possessive in our approach, given a broad understanding of possessivity. The main reason is to differentiate this use from the Patient one, which is illustrated below.

[10] Mind the highlighted option *таблиця* 'table' vs *дуга* 'arc'at the bottom of the screenshot. There is also an option for the universal (in the UD sense), English names of the morphological tags and relations, but we will be using here the translated ones for the convenience of the target Ukrainian user.

Example of the non-possessive use of the pronoun, where it stands for a semantic Patient:

Особливості надання спеціальних дозволів на здійснення господарської діяльності з геологічного вивчення запасів ***газу*** *(метану),* ***його*** *видобування на шахтах визначаються відповідно до цього Закону.* ’The peculiarities of providing special permits for carrying out economic activity on geological exploration of gas (methane), its mining in mines are determined in accordance with this Law.’

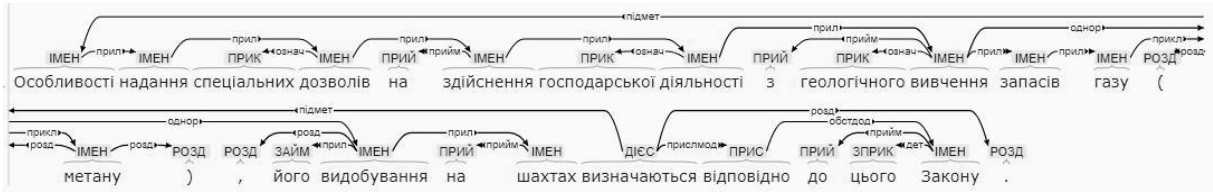


Fig. 4. Parsing results for the Patient pronoun role, tree mode

Even though sentences with several pronouns surrounding the GLF are considered non-grammatical, most likely due to their challenged comprehensibility, they still happen in the real language use. The query [word="їх|її|його"] [lemma=".*(ння|ття)"] [word="їх|її|його"] to Zvidusil treebank returned 373 results. In our case they are good illustrations of a predicate (gerund, in this case) accompanied by both semantic roles, when one of the pronominal uses is possessive (or rather quasi-possessive, as it stands for the Agent role) and the other is personal (stands for the Patient). Of course, many of the hits represent cases when both pronouns refer to different predicates, but there are quite a few examples of the pattern in question, as in the example below:

І ***її*** *розуміння* ***його*** *в ту саму мить стало довершеним.* ‘And at that very moment **her** understanding **of it** was perfect.’

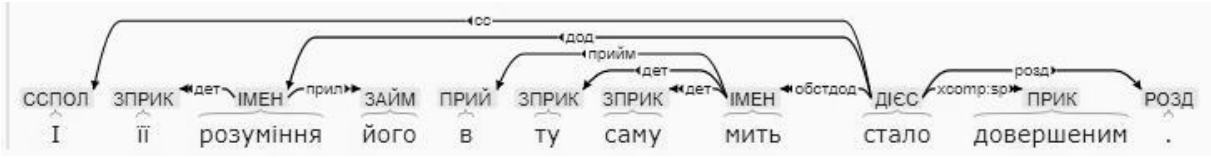


Fig. 5. Parsing results for double pronoun roles, tree mode

We can also notice that when the second pronoun is omitted, the most prominent (and probably the only possible) interpretation of the first pronoun changes to the personal (Patient) one. For more details, comparison with other Slavic languages, and explanation see [11].

Sometimes GLFs lose traces of their origin and function as nouns (for example, they are able to pluralise and can get adjectival modifiers of the non-predicative nature), e.g. *життя, стаття, оголошення*. However, they can still be homonymous to real gerunds, cf. the example below.

… враховуючи те, що обвинувальний ***висновок*** *міститься на майже 500 сторінках у чотирьох томах,* ***його****/adj оголошення може розтягнутися на кілька днів.* ‘Given that the **indictment** takes almost 500 pages in four volumes, **its** announcing may extend for several days.’

Since most frequent use of the word ***оголошення*** ’poster/advertisement/announcement’ is nominal, the parser mistakenly treats the gerundial forms as nominal either. It is only with the help of the semantic hints: activity verb *розтягнутися* ‘extend’ and the temporal modifier of period *на кілька днів* ‘for several days’, that one can identify the gerund proper in this sentence.

The example above clearly belongs to the more complicated cases and is out of the scope of the present research. However, it is a good illustration of the diversity of the problem which proves that we should not expect its full solution using one particular method but should rather seek some approximation. Therefore we will first concentrate on less ambiguous examples, where the lemma of the GLF is sufficient to discern the grammatical quality of the pronoun. Since gerunds most often inherit valency patterns from their ancestor verbs, information about valency of verbs from which the gerunds are derived seems to be a valuable source of information.

Extracting valency information

Valency information was extracted semi automatically from the Dictionary of Ukrainian language (SUM, 26], the biggest currently existing dictionary with definitions of meanings of Ukrainian words), and later revised manually[11]. SUM contains partial grammatical information about entries given in a loose textual form and this was the main source of valency information for us. The digital unparsed version of SUM available at http://sum.in.ua/ was used for this purpose. SUM contains very few noun entries with valency information. Those with reference to verbs do not have any additional grammatical information. A simplifying assumption was taken that gerunds inherit valency patterns of their ancestor verbs. The linking between GLFs and the verbs was established on the basis of semantic definitions of the former which in most cases state explicitly "action according to the meaning of 'verb x'", where 'verb x' is the infinitive form of the gerund's derivative basis. Here is an example of such an entry:

ВИЗНА́ННЯ, я, рідко **ВИЗНАТТЯ́**, я, сер. 1. Дія за значенням визнати 1—3. [http://sum.in.ua/s/Vyznannja].

Gerunds extracted on the basis of word definitions are this way already disambiguated from their nominalised homonyms. For simplicity, the remaining meanings are just not taken into consideration but they exist and add to GLFs' lexical ambiguity. The word *визнання* 'recognition' illustrated earlier has two more lexicalised meanings, and practically all other gerunds do.

Currently the valency database for Ukrainian includes 39207 verb meaning definitions, part of which are mapped to 7450 gerundial forms (gerunds proper). 14454 verb meanings are unambiguously transitive, 20560 are unambiguously intransitive, and 4193 (about 10 %) can be either transitive or not. The derived gerunds comprise 3527 unambiguously transitive, 1407 unambiguously intransitive, and 2516 (about 34 %) ambiguous forms, respectively.

Example of valency presentation and linking:

Для ***його/noun*** *<u>відновлення</u> потрібно десятки мільйонів гривень...* 'For **its** <u>renovation</u> tens of millions hryvnyas are needed…'

The above sample sentence from the experimental test set is linked to two valency dictionary entries. It uses information on transitivity for the ancestor verb through the intermediate dictionary for gerunds, see two corresponding examples of dictionary entries below.

A working valency dictionary entry for a gerund (mind the two possible ancestors indicated here, transitive and intransitive ones):

ВІДНОВЛЕННЯ відновити й відновитися 1-3 verb action:state 1. 3 Дія і стан за знач. відновити й відновитися 1-3.

Working valency dictionary entry for a verb:

ВІДНОВЛЮВАТИ ВІДНОВЛЮВАТИ ВІДНОВЛЯТИ imp. ВІДНОВИТИ перех. 1. S: A:acc -0 S:0 A:acc - acc 0 acc 0 Надавати попереднього вигляду чому-небудь пошкодженому, зіпсованому, зруйнованому; приводити до попереднього стану; поновлювати.

The above pattern includes: the basic form, its phonetic and aspectual variants (*imp*. stands for the 'imperfective'), transitivity marker (*перех*. is short for 'transitive' in Ukrainian), meaning number (1.), lemma related valency pattern, meaning related pattern (in this case these are identical), morphosyntactic pattern for the argument (in this case it is only the direct object expressed by the accusative case), for the lemma and for the meaning. Comparing meaning valency patterns for particular meanings with the general, lemma related ones, gives information about potential ambiguity.

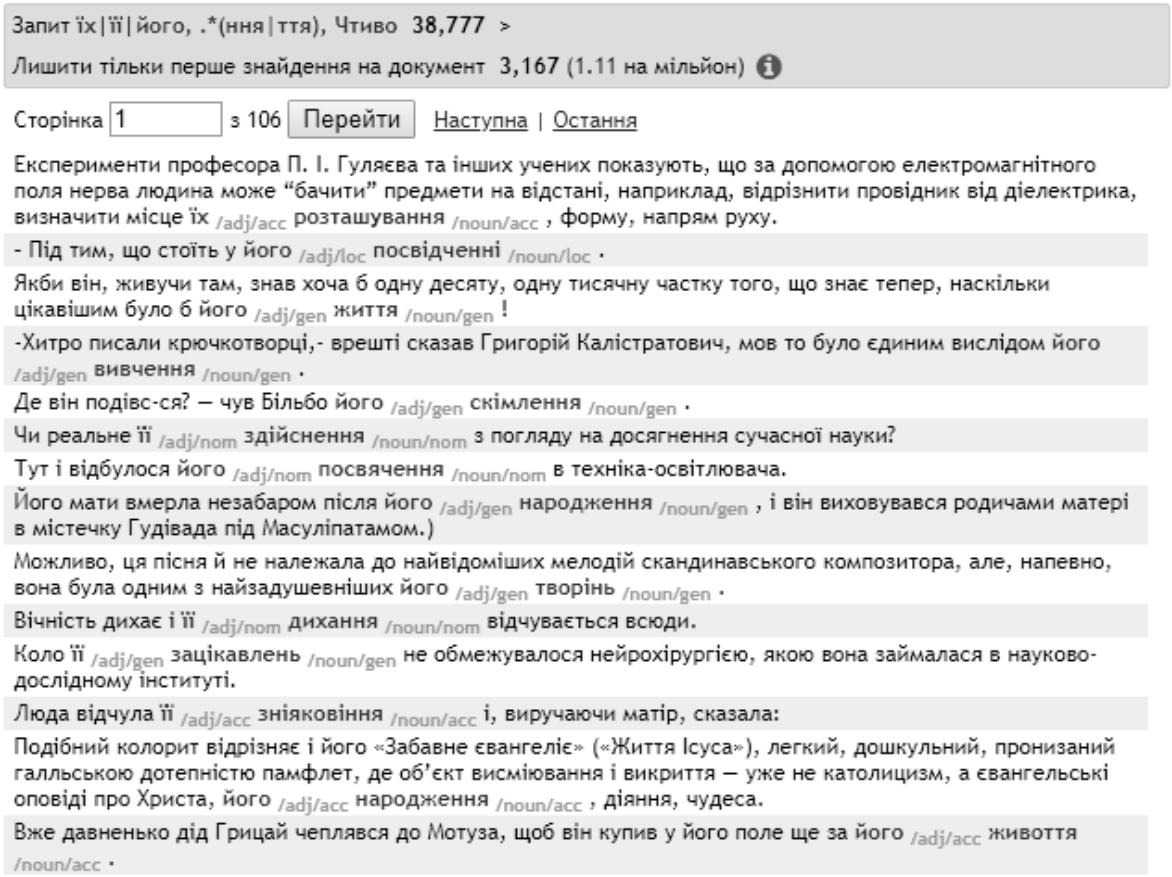
Of course, there are more simplifications and assumptions in the present model to consider: 1) the dictionary does not contain all the existing forms and all the existing meanings of the listed forms; 2) as shown in [21, 23], Ukrainian gerunds are strongly grammatical and can be generated on the fly, making

[11] The presented resource actually grows into an independent valency dictionary. While it is currently in a very raw state, there are plans to enrich it with missing valency information from corpora and publish for machine and human use.

use of the generative power of language [16]. Their presence in a dictionary therefore rather reflects the frequency of their usage and discrepancies with the real language use can be expected. All these circumstances contribute to the grey zone of the experimental results that are responsible for a considerable part of the failures in the experiment. The purpose of the experiment thus was to estimate the scope of this grey zone and of the impact of the valency information in its most basic form for disambiguating GLFs' dependent pronouns. The next section shows some details of this procedure.

4. Experiment

An additional specific test set (SpecTS) with utterances containing a 3rd person pronoun and GLF pattern was created to test the impact of our improving efforts. A simple (string based) query pattern was used to extract them from the Zvidusil Treebank: [word="їх|її|його"] [lemma=".*(ння|ття)"]. The set contains **150** samples from: fiction (60), fora (30), newspapers (30), and additional random genre (30); to randomise the selection only one occurrence per document was allowed. This test set was later added to the testing part of the Gold Treebank to see whether and which of our changes led to corrections of the pronoun labelling by the parser.


Запит їх|її|його, .*(ння|ття), Чтиво 38,777 >

Лишити тільки перше знайдення на документ 3,167 (1.11 на мільйон)

Сторінка 1 з 106 Перейти Наступна | Остання

Експерименти професора П. І. Гуляєва та інших учених показують, що за допомогою електромагнітного поля нерва людина може "бачити" предмети на відстані, наприклад, відрізнити провідник від діелектрика, визначити місце їх /adj/acc розташування /noun/acc , форму, напрям руху.

- Під тим, що стоїть у його /adj/loc посвідченні /noun/loc .

Якби він, живучи там, знав хоча б одну десяту, одну тисячну частку того, що знає тепер, наскільки цікавішим було б його /adj/gen життя /noun/gen !

-Хитро писали крючкотворці,- врешті сказав Григорій Калістратович, мов то було єдиним вислідом його /adj/gen вивчення /noun/gen .

Де він подівс-ся? — чув Більбо його /adj/gen скімлення /noun/gen .

Чи реальне її /adj/nom здійснення /noun/nom з погляду на досягнення сучасної науки?

Тут і відбулося його /adj/nom посвячення /noun/nom в техніка-освітлювача.

Його мати вмерла незабаром після його /adj/gen народження /noun/gen , і він виховувався родичами матері в містечку Гудівада під Масуліпатамом.)

Можливо, ця пісня й не належала до найвідоміших мелодій скандинавського композитора, але, напевно, вона була одним з найзадушевніших його /adj/gen творінь /noun/gen .

Вічність дихає і її /adj/nom дихання /noun/nom відчувається всюди.

Коло її /adj/gen зацікавлень /noun/gen не обмежувалося нейрохірургією, якою вона займалася в науково-дослідному інституті.

Люда відчула її /adj/acc зніяковіння /noun/acc і, виручаючи матір, сказала:

Подібний колорит відрізняє і його «Забавне євангеліє» («Життя Ісуса»), легкий, дошкульний, пронизаний галльською дотепністю памфлет, де об'єкт висміювання і викриття — уже не католицизм, а євангельські оповіді про Христа, його /adj/acc народження /noun/acc , діяння, чудеса.

Вже давненько дід Грицай чеплявся до Мотуза, щоб він купив у його поле ще за його /adj/acc життя /noun/acc .

Fig. 6. Fragment of the concordance on which SpecTS was based

The first step was to modify the tags in the Gold corpus to comply with our revised theoretical assumptions about gerunds described above and retrain the parser. The results improved slightly, see column "fixes" in Table 3 below. The next step was adding basic valency information for GLF, to check how transitivity of the verb from which the GLF was derived influences the parser's behaviour.

The results of the transitivity aware parser were compared against the basic line results. The discrepancies were analysed and grouped. Manually crafted consistency tests[12] were used for automatic

[12] The consistency checks are regenerated upon each new corpus build and can be traced at: https://lab.mova.institute/files/pomylky_robochoho_tb.html.

detection of problematic areas, e.g. pronoun parsed as noun in the context of a gerund derived from an intransitive verb is a signal of a potential error, see Figure 7.

The consistency checks also helped to ensure no errors crept into SpecTS annotations. The last column of Table 3 shows how information on transitivity improves given reasonably consistent (fixed) tagging of pronouns.

The analysis of mistaken parses (i.e. of what the parser “refuses” to learn) in the SpecTS reveals an extended picture of the same problems as detected on the smaller set based on the Gold Treebank, described in greater detail in [11]. However, the bigger scope of the set combined with the transitivity awareness background presents a better overview of the phenomena in question, which makes it easier to define further directions of work to overcome the ambiguity problem. They are presented further in Section 5.

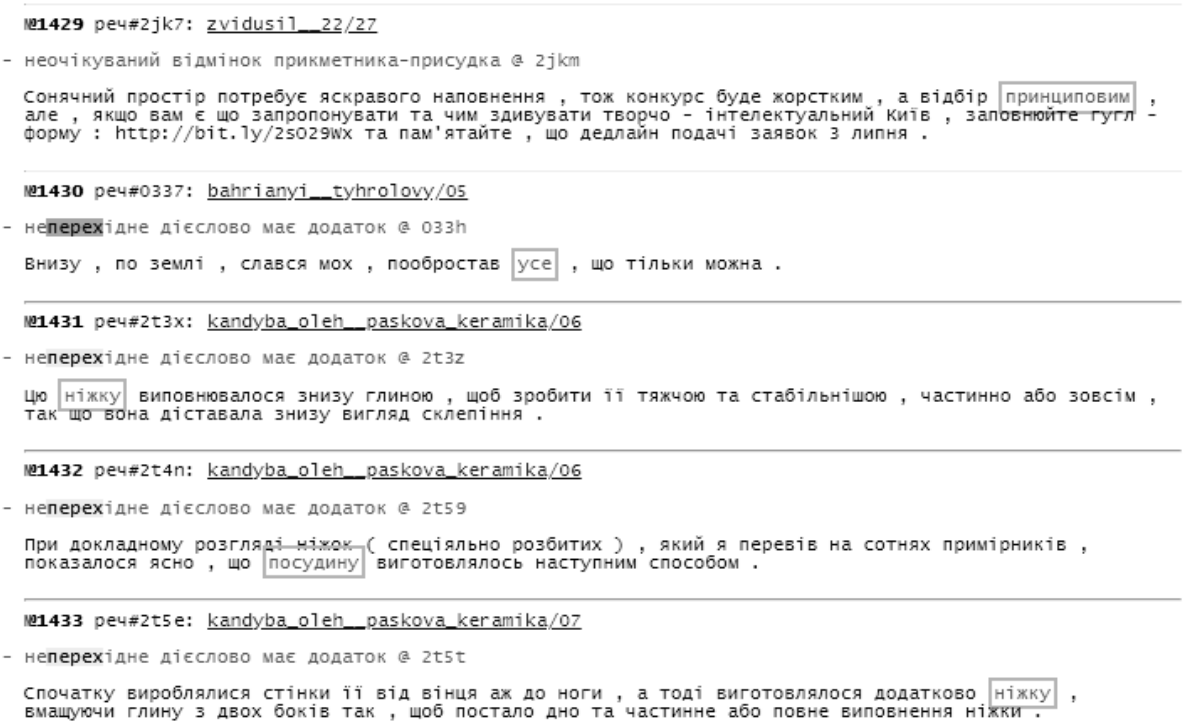


Fig. 7. Fragment of the validation results

Table 3

UDPipe test results on plain text

metric	baseline	fixes	fixes+valency
specific test errs/ %acc	75 (50 %)	71 (**47.33 %**)	66 (44 %)
parts of speech	97.25	97.25	97.27
features	91.48	91.58	91.32
whole tags	90.87	91.03	90.73
lemmas	98.20	98.16	98.24
UAS computed	79.27	**80.11**	79.82
LAS computed	75.52	**76.28**	75.93

Legend (for raws): fixes — model trained on fixed Gold TB. Validation rules were utilising valency dictionary; fixes+valency — valency feature added to the Gold and to the UDPipe's morphological dictionary, UDPipe was retrained; (for columns): computed = morphotagged by UDPipe; the rest of the explanations are the same as for Table 1.

Table 4

UDPipe test results on precomputed (gold) text

metric	baseline	fixes	fixes+valency
features	91.61	91.71	91.32
whole tags	91.00	91.16	90.73
lemmas	98.43	98.40	98.24
UAS	82.10	82.87	79.49
LAS	79.89	80.50	81.90

Legend: precomputed for morphological results (features) and lemmas = pretokenised, for syntactic results (UAS, LAS) = manually premorphotagged.

The "fixes" column shows the improvement of the test results after bringing some consistency in annotating *ŭoɔo, ïï, ïx* in pregerundial position in the Gold standard Treebank. The discrepancies were caused by using different approaches during the manual annotation that were partly caused by diverging guidelines in MTE and UD[13].

Adding valency deserves a longer comment, also because for the whole tags and feature sets the results of the automatic, trained annotation slightly dropped. This is caused by the fact that the transitivity/intransitivity markers for gerunds (partial valency information), after having been fed to the parser, have become an integral additional feature in the tag. Even if all other than transitivity features in the tag were guessed correctly, the failure in one feature means the failure of the whole tag. Therefore, the parser had to guess this marker for all unknown cases as well, including those not listed in the dictionary, and was evaluated accordingly. Let us also remember that the markers were projected from the transitivity values of the verbs from which the gerunds were derived and that there are some naturally explained gaps there.

In many cases the pronoun was mistakenly treated as a separate argument of the preceding verb, so, transitivity of the gerund did not play any role in the interpretation of the pronoun that it followed. If the parser had also been aware of valency demands of all other predicates, verbs in the first turn, the situation might have been different. Some other possible reasons for failures are as follows:

1) The training set was really small as for the needs of machine learning, even though it was good enough to show where the problems can be expected. We need enough training cases for the parser to establish the association between transitivity and the role of the pronoun. At the moment it seems that it relies on the information about particular lemmas rather than on this kind of abstraction.

2) The added training set only included the annotation of the correct pronominal form (DET or PRON) but all other words in the sentences were left untagged. Hence, other parsing mistakes could influence the result of the whole parse.

3) The gerundial context itself is often not enough. It is necessary to include into the training set sentences using the same combination of a pronoun and GLF lexeme where the pronoun has either the possessive or personal interpretation depending on the wider context.

4) The UDPipe parser is based on statistics, which nowadays already starts looking a bit old-fashioned. There are strong indications that neural network based parsers will be able to generate models with better inferring abilities, so that even explicitly fed valency information may be not necessary. However, no matter how good the learning system is, the quality of the data is crucial, so it is better to identify the gaps beforehand.

Disambiguation pointers

The semantic role pointers were ordered from relatively clear cases, where the type of the GLF itself decides whether it can (and need to) take the direct object or not, to more sophisticated ones, with a gradual expansion of the necessary context. We analysed mainly errors but some correctly parsed samples are used for better illustration as well.

[13] It has to be noted that UD's guidelines were much less consistent at the beginning of the project than they are now.

Noun markers:

1) There are many frequently used nouns with **concrete** meaning and no or hardly any connection with the verb which are erroneously tagged as gerund, e.g. *життя* 'life' (stylistically marked), *насіння* 'seed', *місцезнаходження* 'placement', *стаття* 'article', etc. We could list them in a special dictionary and tag in some specific way to make them visible for the tagger as different from other GLFs.

2) **Plural** form can point to nouns, at least we did not find any contradicting examples so far, but this is still to be proved, e.g. *похвалила перші його оповідання* '(she) praised his first short stories'.

3) **Coordination** of the unlike is a known phenomenon in the natural language but in this case is rather marginal, so we may most often expect that if one of the conjuncts can be clearly identified as a gerund or a lexicalisation, then the ambiguous one has the same status. Most of the examples found so far confirm this thesis. At the same time, primary nouns get coordinated more eagerly – only one case was found with gerunds' coordination, the last one from the three listed below. But of course, ambiguous forms also coordinate and then we need to look for other criteria (see examples for the "left head" below).

Все ***її кохання****, ніжні пестощі, піклування про нього і ця її дитяча грайливість - невже оце все тільки удавання?* 'All **her love**, tender caresses, caring for him and this her childish playfulness - is it all just a make-believe?'

Які ***його завдання*** *й обов'язки*? 'What are **his tasks** and responsibilities?' *Густина органічної маси, супутніх порід, рядового вугілля, продуктів* ***їх збагачення*** *і розсортування.* 'Density of organic mass, contiguous rocks, ordinary coal, products of their enrichment and sorting'.

4) Examples of the **left side syntactic head** for pronouns in fact happen quite frequently and deserve more attention, e.g. *позбавити* ***його звання*** *полковника* 'to deprive **him of the rank** of colonel', *озброєння* ***їх знаннями*** *й уміннями* 'arming **them with knowledge** and skills'. The verb on the left expects the direct object on its right side and it is reasonable to let the parser know about this by providing valency information for verbs proper, not only GLFs. Of course, if the verb is modal, then it is more possible that its direct object will be a gerund proper.

Among less stable markers of nouns we can mention **attributes (adjectives)** of non-predicative nature and **verbs with concrete meanings** (also demanding nouns with concrete meanings). Examples are omitted for the sake of sparing space.

Gerund – Agent markers

GLFs derived from **intransitive** only verbs (with no reflexive counterparts), such as: existence/being and their phases; mental processes; social interaction and speech verbs, including independent reflexive and independent reciprocal verbs, will have a quasi-possessive (Agent) argument. To help the parser identify them we can provide more training examples and more specific valency information, i.e. using a dedicated intransitive only tag to differentiate them from lexicalised GLFs (maybe the machine still "considers" action/process more relevant than argument's semantic role while learning)

Existence of **another dependent** of the GLF in the **genitive case** is a good marker of Agent interpretation for the pronoun, e.g. *він не вписується в* ***його розуміння*** *норм моралі суспільства* 'he does not fit into **his understanding** of the norms of society's morals'. It is in fact stronger than the semantic type (Patient/Theme of understanding can be „human" as well as Agent), the genitive expression of Patient/Theme is also possible. On the other hand, for concrete nouns the presence of the genitive is not an obstacle, as the two genitive modifiers belong to different places in the syntactic hierarchy and have different "supersenses" according to (Blodgett, Schneider 2018), see also example about rank of colonel above, for which this test does not work.

Direct object can also be already present in another form, e.g. the **infinitive**, and this also blocks the possibility of Patient/Theme interpretation for the pronoun, e.g.:

Україна також не лишається осторонь цього тренду, популярність блогерів та ***їх вміння*** *доносити інформацію до аудиторії зумовлює появу нових проектів.* 'Ukraine also does not stay away from this trend, the popularity of bloggers and **their ability** to communicate information to the audience leads to the emergence of new projects.'

Gerund – Patient markers

For some verbs direct objects cannot be dropped, i.e., even without the explicitly expressed direct object the only possible reading of the genitive is that of Patient. In other words, if Agent is expressed by a full lexical word in the genitive and no Patient is mentioned at all, the sentence is ungrammatical.

Причини правопорушень — це соціальні явища різного рівня, що призводять до ***їх вчинення*** *на масовому, груповому та індивідуальному рівнях.* 'The causes of offenses are social phenomena of different levels which lead to **their committing** at mass, group, and individual levels.'

Традиційний підхід до виробництва, незалежно від виду продукції — це ***її виготовляння*** *і контроль якості...* 'Traditional approach to a product, regardless of product type - is **its manufacturing** and quality control ...'

In general, if no other pointers are present, nouns are more frequent than gerunds, but this is certainly not a criterion we are looking for. Nouns' contexts are too diverse, and even when described, they may still have several interpretations, which does not allow us to build a robust decision tree. A valency dictionary for nouns could be a good starting point in this direction.

Anaphora

Anaphora is a costly tool and should be used when other resources are exhausted. Gender and number markers are often helpful to identify it (although see less standard example below). In most cases the principle of the closest mentioned entity can be used. Here are some examples which demonstrate this:

Природні катаклізми були, є і будуть, тому в ***їх передбаченні*** *чуда не бачу.* 'Natural disasters did, do, and will happen, therefore, I do not see any miracle in **their foreseeing**.'

Голодомор це злочин проти української нації, а отже ***його заперечення*** *це приниження гідності й нації також.* 'The Holodomor is a crime against the Ukrainian nation, and therefore **its denial** is the humiliation of dignity of the nation as well.'

– Хитро писали крючкотворці,- врешті сказав Григорій Калістратович, мов то було єдиним вислідом ***його вивчення***. '"The crooked writers wrote cunningly", – finally said Hryhoriy Kalistratovych, as if that was the only result of **his study**.'

… ініціатором "воєнних дій" у торгівлі був уряд України, який почав ***її запровадженням*** *ПДВ на російські товари.* '… the "military action" in trade was initiated by the government of Ukraine which began **it** by **introducing** VAT on Russian goods.'

The last example is not trivial. It may be solved by the presence of the explicit object in the genitive case but it also has an interesting case anaphora, where the plural "military actions" are referred to by the singular feminine pronoun *її* 'her' by replacing them on the fly with *війна* 'the war'. This may complicate anaphora resolution process if it is approached too mechanically.

Cases which remain ambiguous

Filtering off **reflexive** uses most often is connected with determining the importance and relevance of Agent, which is a semantic task. Even if Agent is present it may be more important to accentuate that something happened to the object and then the reflexive form is the basis of GLF derivation. Such cases remain ambiguous for human annotators as well, their proper interpretation can be provided only by the authors of the utterances.

Вітаю батька Теревенів з ***його народженням:)*** 'My congrats to the father of Tereveni with <**his**|**its**> **birth** :)' (Tereveni, lit. 'chitter-chatter' or 'chat', is the title of a forum portal; probably "birthday" was meant instead of "birth", although the latter use is acceptable in the colloquial language.)

Його мати вмерла незабаром після **його народження** 'His mother died soon after <**his birth**|**giving birth to him**>.'

In the first case we certainly deal with reflexivisation because a male cannot give birth. (Another problem is that "birth" can be used metaphorically here and refers to "birth" of the forum in par with "father of the forum", but this is beyond our concern at the moment). The last example can have either interpretation.

Reflexivisation is just one of numerous illustrations of underspecifying in the language. Many other ambiguous constructions may be resolved with the help of a bigger context and human help. They are subject to further investigation.

Concluding remarks and future work

A detailed analysis of the selected pronouns ambiguity problem revealed that many different layers of language representation are involved and that in the end this is not the simplest problem to start with when trying to improve the parsing accuracy. However, the choice of the subject matter was dictated by practical needs, and, when one starts exploring a new area, there is always a risk that the path may be more complicated than one expected.

In [11] we had defined the principles of consistent annotation of pronouns in one specific, frequently used construction and corrected the Gold Treebank annotations according to them. In this follow-up research we have tested the parser's performance after making it aware of the transitivity status of the gerund's derivational basis verb. This experience shows that bringing consistency to data is as much important as feeding additional information to the parser. Machine learning works according to the principle "As you sow, so shall you reap" and very often revising the structure of the fed input, better organisation of the internal logics, helps the machine to grasp the patterns behind the data. This also works in the opposite direction – if the machine cannot learn the pattern (and the algorithm works well for other data) one of the reasons may be that the data are not well organised.

No matter who does the tagging task, the machine or the human, they need to have good instructions in the form of a theoretical description (for humans) or consistently tagged training data in accordance to the theory (for machines), otherwise failures are going to be propagated in either case. To prepare this background we need to understand the subject matter profoundly, which basically means that we have to solve the problem fully at the theoretical level to be able to supervise the machine learning process. Sufficient training data means that for each potentially ambiguous situation we need to prepare a training set which will make it possible for the machine to induce the underlying rules. The experience with the presented research problem also teaches us that quality prevails over quantity, and working with the natural language is not only (if any) the matter of the big data collection. Corpus gives the researchers an excellent opportunity to verify, develop, and tune their theories. Bootstrapping with specific case training sets seems a very promising corpora correction technique.

The directions for future work are outlined in Section 5 and we may continue our investigation by moving from the "lower hanging fruit" towards the "higher hanging" ones. Besides, we may consider studying the behaviour of other suffix groups of deverbal nouns, the question of valency patterns inheritance, impact of the semantic categories of verbs and their lexical and grammatical aspect, creating bigger and more specific training sets.

And a final remark: while preparing English translations of the Ukrainian examples for this paper we were using the Google Translate service. The quality of machine translation has recently advanced significantly and is no longer the subject of users' jokes but our construction was very often translated with errors, which means that anaphora resolution, valency patterns, and semantic roles assignment are still weak points for the artificial intelligence. For us this also means that this study and its possible continuation, by contributing to the linguistic knowledge in general, can be beneficial for other natural language processing tasks.

Literature

1. Blodgett, A., Schneider, N. (2018). Semantic Supersenses for English Possessives. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki (Japan).

2. Danielewiczowa M. (2017). Polskie nazwy czynności i wytworów czynności w świetle walencji motywujących je czasowników /Polish Action Nominals in the Light of the Valency of the Corresponding Verbs. *Prace Filologiczne*, tom LXX, p. 143–157.

3. de Marneffe M.-C., Dozat T., Silveira N., Haverinen K., Ginter F., Nivre J., and Manning Ch.-D. (2014). Universal Stanford Dependencies: A cross-linguistic typology. *LREC.*

4. Dozat, T., Qi, P., Manning Ch.-D. (2017). Stanford's Graph-based Neural Dependency Parser at the CoNLL 2017 Shared Task. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, Canada, August 3–4, 2017, p. 20–30.

5. Erjavec T. (2009). MULTEXT-East Morphosyntactic Specifications: Towards Version 4. *Proc. of the MONDILEX Third Open Workshop*, Bratislava, Slovakia, 15–16 April, 2009.

6. Kocková, J. (2017). Substantiva mezi slovesem a jménem Substantiva na -ní (-tí) / -ние (-тие) v češtině a ruštině ve světle paralelního korpusu. *Časopis pro Moderní Filologii* 99, Č. 1, p. 55–64.

7. Kotsyba N. (2013). Overview of the Ukrainian language resources within the multilingual European MULTEXT-East project, v. 4. *Вісник Національного університету "Львівська політехніка". № 770: Інформаційні системи та мережі*. p. 122–129. http://science.lp.edu.ua/sisn/vol-770-no-2013-1

8. Kotsyba, N. (2014). How light are aspectual meanings?: A study of the relation between light verbs and lexical aspects in Ukrainian. Robering, K. (ed.) *Events, Arguments, and Aspects. Topics in the Semantics of Verbs*. Studies in Language Companion Series, vol. 152, pp. 261–300.

9. Kotsyba N. (2014). Using Polish Wordnet for Predicting Semantic Roles for the Valency Dictionary of Polish Verbs. Przepiórkowski A., Ogrodniczuk M. (eds) *Advances in Natural Language Processing. NLP 2014. Lecture Notes in Computer Science*, vol 8686. Springer International Publishing Switzerland, p. 202–207. https://link.springer.com/chapter/10.1007/978-3-319-10888-9_21

10. Kotsyba, N., Moskalevskyi, B. (2018). An essential infrastructure of Ukrainian language resources and its possible applications. *SlaviCorp 2018, 24–26 September 2018, Charles University, Prague, Book of Abstracts*. https://slavicorp.ff.cuni.cz/wp-content/uploads/sites/144/2018/09/SlaviCorp2018_Book_of_Abstracts.pdf

11. Kotsyba, N., Moskalevskyi, B. (2018). Syntactic and morphological ambiguity of the deverbal nouns' arguments in Ukrainian and ways of its resolution. *Prace Filologiczne*, vol. VXXII, Warsaw, p. 193–210.

12. Levin, B. and Rappaport Hovav M. (2005). *Argument realization*. Cambridge: Cambridge University Press.

13. Panevová, J. (2017). Od valence slovesa k valenci substantiv a adjektiv/From Valency of Verbs to Valency of Nouns and Adjectives. *Prace Filologiczne*, vol. LXX, Warsaw, p. 59–72.

14. Pazelskaya, A. (2007). Argument structure in Russian deverbal nouns in -nie. *Studies in Formal Slavic Linguistics*, ed. Franc Maršič and Rok Zeucer, p. 255–272. Peter Lang.

15. Petrov S., Das D., and McDonald R. (2012). A universal part-of-speech tagset. *LREC*.

16. Pustejovsky, J. (1995). *The Generative Lexicon*, MIT Press, Cambridge, MA.

17. Straka M., Hajič J., Straková J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, May 2016.

18. Zeman, D. (2008). Reusable Tagset Conversion Using Tagset Drivers. *LREC*.

19. Vykhovanets I., Horodenska K. *Theoretical Morphology of Ukrainian Language: Academic Grammar of Ukr. Lang*. Kyiv: Pulsary, 2004. [Ukrainian]

20. Kobozeva I.M. About Possessivity in Russian: Possessive Predicates and the Genitive. *Acta Linguistica Petropolitana. Scientific Papers of Institute for Linguistic Research RAS*. T. XI. P. 1. Categories of Noun and Verb in the System of Functional Grammar. Nauka, S. Petersburg, p. 249–271, 2015. [Russian]

21. Kurylo, O. *Considerations about the Modern Ukrainian Literary Language*. Solomiya Pavlychko's Publishing House "Osnovy", Kyiv, 2004 (reprint from Knyhospilka, 1925). [Ukrainian]

22. Pazelskaya A. G., Tatevosov S. G., The Deverbal Noun and the Structure of the Russian Verb. V. A. Plungian, S. G. Tatevosov (ed.), *Research on Verbal Derivation. Languages of the Slavic Culture*. Moscow, p. 348–380, 2008. [Russian]

23. Pchelintseva, J. E. The Grammatical Status and Aspectuality of Deverbal Nouns of Action in Ukrainian (on the background of Russian and Polish). *Izvestiya VGPU. Philological Studie*s. Volgograd, 2015. [Russian]

24. Syniavskyi O. N. *The Norms of the Ukrainian Literary Language*. Ukrainian Publisher, 2nd edition, Lviv, 1941. [Ukrainian]

25. *Syntactic corpus search interface*. Retrieved March 19, 2019, from http://www.mova.info/syntaxis_search.aspx. [Ukrainian]

26. *SUM – Dictionary of Ukrainian language in 11 volumes.* „Naukova Dumka", Kyiv, 1970–1980. Digital version of SUM. Retrieved March 19, 2019, from http://sum.in.ua/. [Ukrainian]

27. *IU Gold – Syntactic Corpus of the Ukrainian Language* (Gold Standard Treebank of Ukrainian). Institute for Ukrainian, NGO, 2018. Retrieved March 19, 2019, from https://mova.institute [Ukrainian]