УДК 81’322

***Алла Міщенко***
***Кіровоградський державний педагогічний університет імені Володимира Винниченка***

# СТВОРЕННЯ ПАРАЛЕЛЬНОГО БАНКУ ДЕРЕВ ДЛЯ НІМЕЦЬКОЇ ТА УКРАЇНСЬКОЇ МОВ

*Описано створення паралельного банку дерев для української та німецької мов. Було здійснено ручне тегування й лематизацію токенів для української мови, створено набір тегів для анотування українських речень на рівні синтаксичної структури; проведено вирівнювання й визначення повних або часткових відповідників на рівні як термінальних, так і нетермінальних символів. Для анотування банку дерев українською мовою було застосовано формат TIGER-XML, адаптований до потреб формального опису граматичної структури української мови.*

***Ключові слова:*** банки дерев, анотація, тегування частин мови, вирівнювання, переклад.

Інновації в галузі економіки, політики й технологій суттєво впливають на зростання попиту на перекладацькі послуги, а глобалізаційні процеси й формування єдиного полілінгвокультурного суспільства обумовлюють необхідність перекладу контенту багатьма мовами, що супроводжується поступовим, але постійним зростанням обсягів перекладу, збільшенням кількості мов перекладу та витрат на нього. З огляду на це необхідно шукати можливостей оптимізації процесу перекладу. Таким “соломоновим рішенням” стала концепція “пам’яті перекладу”, яка ґрунтується на повторному використанні раніше перекладених сегментів контенту. Згідно з визначенням, пам’ять перекладу – це “архів багатомовних сегментованих, вирівняних (aligned), проаналізованих (parsed) та класифікованих текстів <…>, який дозволяє зберігати попередньо вирівняні сегменти текстів й здійснювати їх пошук відповідно до заданих параметрів” [Eagles : 140].

Концепція пам’яті перекладу вперше була реалізована на програмному рівні ще у 60-ті рр. XX ст. Перший її прототип було створено для Європейської спілки вугілля та сталі і використовувався для перекладу багатомовного контенту галузевих текстів.

Ефективність методу "повторного застосування", з одного боку, та обсяги перекладу, які постійно зростали, з іншого боку, значно прискорили процес створення й розбудови програмного забезпечення для оптимізації процесу перекладу; а можливості збереження контенту в електронній формі обумовили створення лінгвістичних ресурсів як для академічного, так і для комерційного застосування.

Асортимент програмних продуктів (Tools), які вможливлюють повторне використання раніше створеного багатомовного контенту, значно розширився. На сучасному етапі він представлений такими інструментами: системи пам'яті перекладу (Translation Memory Systeme, TMS), системи управління контентом (Content Management Systeme, CMS), системи підтримки технічного автора (Authoring Systeme, AS) та ін., які детально охарактеризовані в науковій літературі [Massion 1995; Reinke 2004; Boiko 2005; Closs 2011; Drewer / Ziegler 2011] та технічній документації на CLAT (Controlled Language Authoring Technology) (2010).

У статті описано експеримент зі створення банку дерев (Treebanks) для української мови (Ukrainian language) та його вирівнювання (Alignment) з банком дерев німецької мови. Робота виконувалася під керівництвом проф. О. Капанадзе і була б неможлива без його консультацій, підтримки та редагування.

Паралельні банки дерев розглядаються сьогодні як важливі ресурси для навчання перекладу, а також для вирішення окремих прикладних завдань у галузі комп'ютерної лінгвістики. У процесі навчання банки дерев слугують для унаочнення й ілюстрації контрастивних феноменів мовних пар. У корпусній лінгвістиці вони використовуються як ефективні ресурси для дослідження й аналізу синтаксичної структури мовних пар. В інших галузях комп'ютерної лінгвістики паралельні банки дерев застосовуються, зокрема, для тренування й оцінювання ефективності парсерів. У перспективі паралельні банки дерев після конвертування у стандартизований формат пам'яті пере-

кладу можуть імпортуватися в комерційні системи пам'яті перекладу на кшталт SDL, Across, DejaVu, MemoQ та застосовуватися в них у процесі перекладу.

Для проведення експерименту було відібрано 40 речень із лексикону валентності для NLP німецькою мовою, укладеного в рамках проекту GREG (German-Russian-English-Georgian) [Kapanadze O, Wanner L., Klatt S 2002 *:* 11]. Ці речення перекладені українською мовою і вручну анотовані з урахуванням особливостей української мови.

У процесі створення паралельного банку дерев, що вирівнюється автоматично, ключове значення відіграють програми синтаксичного аналізу (Parser), які на етапі генерування монолінгвальних банків дерев аналізують синтаксичну структуру речень, оскільки для української мови ще не створено ресурсів для автоматизованого лінгвістичного анотування речень із можливістю їх наступної візуалізації у Synphaty й подальшого застосування для побудови паралельних банків дерев. Таким чином, автоматичне створення паралельного банку дерев для німецької та української мов було неможливим, але привабливість для перекладацької галузі ідеї автоматичного генерування банків дерев із паралельних текстів з їх наступною конвертацією у пам'ять перекладу спонукала нас перевірити можливість опрацювання мов із кириличним шрифтом програмними продуктами Synphaty і TreeAligner, розробленими спеціально для побудови й візуалізації паралельних банків дерев для мов з латинським шрифтом, а також потенційні можливості побудови паралельних банків дерев для української та німецької мов з огляду на їхні морфологічні й структурні дивергенції.

Досягнення запланованої мети передбачало вирішення таких завдань:

- створення й візуалізацію монолінгвального банку дерев для української мови;
- конвертування даних, виведених у форматі tig (Synpathy), у формат xml (TreeAligner), вирівнювання (Alignierung) й ві-

зуалізацію німецько-українського банку дерев у програмі TreeAligner з подальшим редагуванням вручну.

На нашу думку, інструменти, використані для генерування й візуалізації паралельного банку дерев, потребують детальнішого пояснення.

Для вирівнювання паралельних речень використано програму Stockholm TreeAligner. Ця програма виводить дані у форматі xml і, таким чином, дозволяє підготувати паралельні речення до експорту в системи пам'яті перекладу. Але на першому етапі речення для кожної мови лінгвістично анотуються й візуалізуються у формі графів у програмі Synphaty.

Synphaty розроблена в інституті психолінгвістики Макса Планка в голландському місті Неймеген. Основним компонентом програми є візуалізатор синтаксичної структури речення Syntax-Viewer, розроблений в інституті машинної обробки природної мови у Штутгарті (Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart) для дослідницького проекту TIGER. Synphaty підтримується операційними системами Windows, MAC OS та Linux, а ліцензія на її застосування з академічною метою безкоштовна. (Детальніша інформація про Synphaty розміщена на сайті інституту психолінгвістики Макса Планка, http://www.mpi.nl/tools/synpathy.html).

Структура формату tig (Synphaty) складається з двох частин: заголовка (Header) та основної частини (Body). Заголовок містить метадані: назва корпусу, дата його укладання, укладач, експлікація використаних тегів тощо. Основна частина складається з: 1) ідентифікації графа (S ID); 2) початкового символа дерева (S); 3) термінальних вузлів (Terminals, рівень слів); 4) не термінальних вузлів (Nonterminals, рівень фраз: NK, NN та ін.); 5) первинних (SB, HD, OA) та вторинних ребер (NK, NN), які пов'язують термінальні й нетермінальні вузли графа та експлікують їхні синтаксичні функції. Крім того, у графі виводяться характеристики термінальних вузлів (terminal node features: PPER, VVFIN, ADJA, NN) та їхні граматичні значення (Mask. 3. Sg. Nom.; 3. Pl.Pres. Ind.; Akk.Pl.).

Для української мови анотування здійснювалося вручну в програмі Notepad++, а потім результати візуалізувалися у вигляді дерева із застосуванням програми Synphaty.

Як для Synphaty, так і для TreeAligner застосовувався набір тегів, спеціально створений для генерування банку дерев корпусу NEGRA (Stuttgart-Tübinger Tagset, STTS, G. Smith, 2003), який доповнено необхідними для анотування українських речень тегами. Він містить теги для анотування одинадцяти класів слів (Part of Speech, POS): Verbs V, Nouns N, Adverbs ADV, Conjunctions KO, Articles ART, Adpositions AP, Adjectives ADJ, Interjections IT, Pronouns P, Particles PTK, Cardinal Numbers CARD.

На рівні термінальних вузлів цим класам присвоюються відповідні характеристики (features) (табл. 1).

***Таблиця 1.*** **Характеристики дієслова**

| Клас | Типи дієслова | Форми дієслова | Характеристики дієслів на рівні термінальних вузлів |
|---|---|---|---|
| **V** | **A** ‘Auxiliar ’ | **FIN** ‘finit ’<br>**INF** ‘infinit ’<br>**IMP** ‘imperativ ’<br>**PP** ‘Partizip Perfekt ’ | **VAINF** ‘have ’<br><br>**VAIMP** ‘be ’<br>**VAPP** ‘had ’ |
| | **M** ‘Modal ’ | **FIN** ‘finit ’<br>**INF** ‘infinit ’<br>**PP** ‘Partizip Perfekt ’ | **VMFIN** ‘könnte ’<br>**VMINF** ‘können ’<br>**VMPP** ‘gekonnt ’ |
| | **F** ‘Full ’ | **FIN** ‘finit ’<br>**INF** ‘infinit ’<br>**IZU** ‘Infinitiv mit zu ’<br>**IMP** ‘imperativ ’<br>**PP** ‘Partizip Perfekt ’ | **VFFIN** ‘gebt … ab ’<br>**VFINF** ‘abgeben ’<br>**VFIZU** ‘abzugeben ’<br>**VFIMP** ‘gib … zu ’<br>**VFPP** ‘abzugegeben ’ |

Характеристики термінальних вузлів експлікуються граматичними значеннями, яких вони можуть набувати у реченні. У табл. 2 наведено граматичні категорії, характеристики термінальних вузлів, які можуть мати ці граматичні категорії, і ті граматичні значення (value), які можуть присвоюватися характеристикам термінальних вузлів.

***Таблиця 2.*** **Елементи анотації для термінальних вузлів**

| Грам. категорія | Характеристики термінальних вузлів | Значення характеристик термін. вузлів |
|---|---|---|
| Genus | ADJA, ART, APPRART, NE, NN, PDS, PDAT, PIAT, PIS (teilweise), PPER, PPOSAT, PPOSS, PRELS, PRELAT, PWAT, PWS | Masc, Fem, Neut |
| Kasus | ADJA, ART, APPRART, NE, NN, PDAT, PDS, PIAT, PIS, PPER, PPOSAT, PPOSS, PRELS, PRELAT, PRF, PWAT, PWS | Nom, Gen, Dat, Acc. |
| Numerus | ADJ, ART, APPRART, NE, NN, PDAT, PDS, PIAT, PIS, PPER, PPOSAT, PPOSS, PRELS, PRELAT, PRF, PWAT, PWS, V.FIN, V.IMP | Sg, Pl |
| Grad | ADJA, ADJD | Pos, Comp, Sup |
| Person | VVFIN, VAFIN, VMFIN, PPER, PRF | 1, 2, 3 |
| Tempus | VVFIN, VAFIN, VMFIN | Pres, Past |
| Modus | VVFIN, VAFIN, VMFIN | Ind, Subj |
| Nichtfinitheit | VVINF, VAINF, VMINF, VVPP, VAPP, VMPP, VVIMP, VAIMP, VVIZU | Inf, Psp, Imp, Infzu |

Графічний візуалізатор унаочнює синтаксичну структуру речення у формі графа й дозволяє редагувати його вручну (“rebuildung” und “re-tagging”) як для термінальних, так і для нетермінальних символів. Остаточне опрацювання, наприклад термінальних символів, здійснюється у вікні управління термінальними символами. Тут можна змінювати назву термінальних вузлів, додавати нові або видаляти існуючі термінальні вузли тощо. Результат редагування зберігається й може експортуватись у формі tig-файлу чи графа за необхідності.

На другому етапі файли формату tig (Synphaty-Output) конвертувались у формат xml (TreeAlingner-Input), у якому зберігались анотовані еквівалентні речення німецькою та українською мовами і вирівнювались вручну у Notepad++, після чого візуалізувалися й редагувались у програмі Stockholm TreeAligner. Приклад вирівнювання паралельних речень у формі графа ілюструє рис 1.

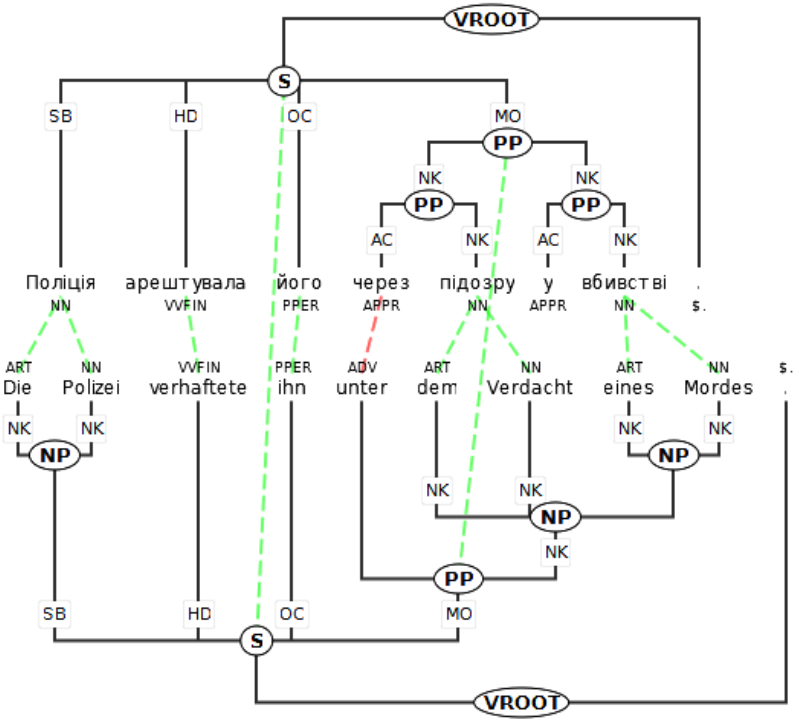


**Рис. 1. Візуалізація білінгвального дерева у TreeAligner**

Паралельні дерева дозволяють унаочнювати дивергенції у структурі мов. Охарактеризуємо окремі з них для німецької та української мов.

Залежно від типу речення нормативна граматика німецької мови чітко визначає порядок слів у реченні й позиціонування підмета, додатка, присудка та частин складеного присудка. На відміну від регламентованої структури німецького речення, порядок слів у реченнях української мови відносно вільний і визначається, насамперед, функціональною перспективою речення, яка обґрунтовується:

- теорією мовленнєвих актів, яка ґрунтується на гіпотезі, що структура повідомлення – це мовленнєва дія, детермінована ілокуцією автора [Austin 2005 : 17];
- темо-ремною теорією, яка відстоює гіпотезу про те, що лінійна послідовність структурних елементів повідомлення підпорядковується перебігу думок людини у напрямку від відомого (тема) до нового (рема) й визначається такими чинниками: комунікативна ситуація, контекст, ставлення автора до потенційного реципієнта, що є необхідною передумовою для забезпечення процесу комунікації [Lutz 1981 : 12].

Рід іменників у німецькій мові визначається граматичними маркерами і/або артиклями. Артикль як граматична категорія відсутній в українській мові, а його функції (ідентифікації, індивідуалізації та генералізації іменника) передаються лексичними (займенниками) або граматичними засобами (формотворчими морфемами).

Доповнення з прийменником у давальному чи знахідному відмінках, які слугують у німецькій мові для позначення напрямку чи інструмента, подекуди передаються українською мовою прямими або непрямими додатками з прийменниками в орудному чи місцевому відмінках.

Німецькі речення характеризуються трьома типовими ознаками:

- двочленність – обов'язкова присутність двох головних членів речення: підмета та присудка;
- вербальний характер речення: частиною присудка, також складеного іменного присудка, завжди виступає дієслово або дієслово-зв'язка;
- чітко визначене позиціонування присудка у реченні залежно від його типу: розповідне, питальне, спонукальне.

Жодна із зазначених ознак не типова для української мови. Тому присудок в українській мові може випускатись або мати структуру, відмінну від структури німецького присудка:

укр. *Відлига.* – нім. '*Es taut.'* замість *'*Taut.'*

укр. *Це стіл.* – нім. *'Das ist ein Tisch.*' замість *'*Dies Tisch.'*

укр. *Хворіти.* – нім. '*krank' sein.*

Український присудок у формі повнозначного дієслова німецькою мовою може відтворюватися номінальною групою з частково десемантизованим дієсловом, напр.:

укр. *об'єднуються* – нім. '*Verbindungen eingehen'*

укр. *ризикують* – нім. '*Risiken eingehen'*

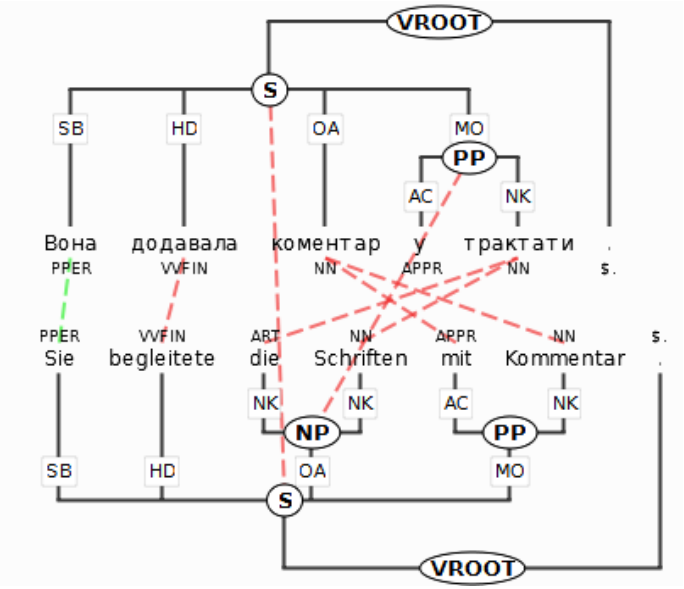


**Рис. 2. Дивергенції на рівні заперечення**

Присвійні займенники у німецькій мові слугують засобом вираження належності, упорядкування або єднання. Вони ставляться перед іменником і відповідають на питання *чий*, *чия*, *чиє*?

Кожному особовому займенникові у німецькій мові відповідає присвійний займенник, який також узгоджується у роді, числі та відмінку із постопозиційним іменником. Присвійні займенники української мови не мають відповідників серед особових займенників й функціонують як омоніми: укр. *свій* – нім. '*mein'*, '*dein'*, '*sein'* і под.

У заперечних реченнях у німецькій мові використовуються такі лексичні засоби:

- частки *nicht;*
- займенники *kein, keiner, niemand, nichts;*
- прислівники *nirgends, niemals, nie, nimmer, nirgendwo;*
- парні сполучники *weder...noch;*
- префікси: *un-, miss-, a-, il-* u.a.;
- еквівалентні речення *nein, doch.*

В українській мові існують усі відповідники німецьких заперечень за винятком заперечного займенника *kein* та його варіантів, які вживаються перед іменниками. Проте подвійне заперечення, типове для українських речень, виключає нормативна граматика німецької мови: укр. *ніхто не* – нім. *niemand.*

Подекуди конвенціоналізовані мовленнєві зразки настільки відрізняються, що їх неможливо перекладати еквівалентними лексичними засобами, а відтворення семантики речення мовою оригіналу вимагає уживання автентичних засобів мовою перекладу.

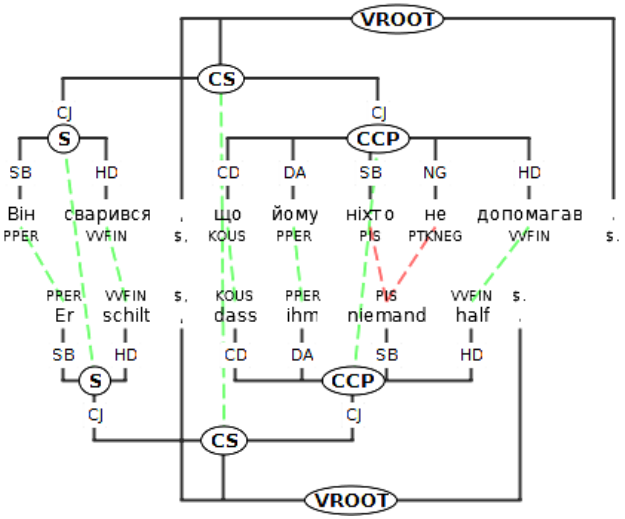


**Рис. 3. Дивергенції на рівні прагматики**

Проте дивергенції на морфологічному, синтаксичому й прагматичному рівнях німецької та української мов уможливлюють застосування вище зазначених інструментів для генерування паралельного банку дерев для цієї пари мов.

Створення таких лінгвістичних ресурсів має вагоме значення у процесі навчання студентів, для проведення наукових досліджень та прикладного застосуваня. У навчальному процесі вони слугують наочним матеріалом для демонстрації дивергенцій та конвергенцій контрастивної пари мов. У перекладацькій галузі такі ресурси мають значну перевагу порівняно з комерційними системами пам'яті перекладу, які створюються переважно на основі статистичних методів. Інтеграція інтелектуальних лінгвістичних ресурсів такого типу у системи пам'яті перекладу та системи машинного перекладу забезпечує якісний автоматичний переклад контенту.

1. *Austin John L.* How to do things with words / John L. Austin. – Cambridge, 2005. 2. *Boiko B.* Content management bible / B. Boiko. – Indianapolis, 2005. 3. *Drewer P.* Technische Dokumentation : eine Einführung in die übersetzungsgerechte Texterstellung und in das Content-Management / P. Drewer, W. Ziegler. – Würzburg, 2011. 4. CLAT-Client-Manual. – Saarbrücken, 2010. 5. CLAT-Intro. – Saarbrücken, 2010. 6. CLAT-In-For-Word-Manual. – Saarbrücken, 2010. 7. CLAT-UMMT-Manual-EN. – Saarbrücken, 2010. 8. CLAT-UMMT-Manual-DE. – Saarbrücken, 2010. 9. *Closs S.* Single Source Publishing : Modularer Content für EPUB & Co / S. Closs. – Frankfurt a. M., 2011. 10. Eagles. Evaluation of natural language processing systems. Final report (First phase). – Access mode : ftp://issco-ftp.unige.ch/pub/ewg96.pz.gz [Zugriff: 23.12.2002, 22:20 MEZ]. 11. *Kapanadze O.* Towards a semantically motivated organization of a valency lexicon for natural language processing: A GREG Proposal / O. Kapanadze, L. Wanner, S. Klatt // Proceedings of the EURALEX conference, Copenhagen, 2002. 12. *Kapanadze O.* Verbal Valency in Multilingual Lexica / O. Kapanadze // Workshop abstracts of the 7th language resources and evaluation conference-LREC2010. – Valletta, 2010. 13. *Lutz L.* Zum Thema "Thema" : Einführung in die Thema-Rhema-Theorie / L. Lutz. – Hamburg, 1981. 14. *Massion F.* Translation Memory Systeme im Vergleich / F. Massion. – Reutlingen, 2005. 15. *Reinke U.* Translation Memories: Systeme – Konzepte – linguistische Optimierung / U. Reinke // Fachrichtung Angewandter Sprachwissenschaft sowie Übersetzen und Dolmetschen der Universität des Saarlandes. – Sabest : Saarbrücker Beiträge zur Sprache- und Translationswissenschaft. Bd. 2. – Frankf./M.; Berlin [u.a.], 2004. 16. *Samuelsson Y.*

Presentation and representation of parallel tree-banks / Y. Samuelsson, M. Volk // In Proceedings of the Treebank-Workshop at Nodalida, Joensuu, 2005. 17. *Samuelsson Y.* Phrase alignment in parallel treebanks / Y. Samuelsson, M. Volk // In Proceedings of 5th Workshop on Treebanks and Linguistic Theories, – Prague, 2006. 18. *Searle J.* Speech acts: an essay in the philosophy of language / J. Searle. – Cambridge [u.a.], 2005. 19. *Smith G.* (2003), A Brief Introduction to the TIGER Treebank, Version 1. / G. Smith. – Potsdam, 2003. 20. Synphaty: Syntax Editor. – Manual. – Nijmegen: Max Planck Institute for Psycholinguistics, 2006.

*Описано создание параллельного банка деревьев для украинского и немецкого языков. Проведено ручное тегирование и лемматизацию токенов для украинского языка; создан набор тегов для аннотирования украинского языка на уровне синтаксической структуры; выровнены и определены полные или частичные совпадения как для терминальных, так и для нетерминальных символов. Для аннотирования банка украинских предложений использован набор тегов в формате TIGER-XML, адаптированный к потребностям формального описания грамматической структуры украинского языка.*

***Ключевые слова:*** банки деревьев, аннотация, тегирование частей речи, выравнивание, перевод.

*In this paper, we describe outcomes of an experiment on building a parallel Treebank for bridging the Ukrainian language with the German language. The aim of the mentioned experiment was: manually tagging and lemmatization of tokens for Ukrainian corpora; establishing of the compatible tagset for Ukrainian and introduction of the specific syntactic phrasal categories; production of the parallel trees from the bilingual resources; alignment of the German-Ukrainian parallel trees; determining "good" and "fuzzy" matches between the non-terminal and terminal nodes across the syntactic structures of the languages involved. The Ukrainian Treebank was annotated according to an adapted version of the German TIGER guidelines with the necessary changes relevant to the Ukrainian grammar formal description.*

***Keywords:*** treebanks, annotation, POS, alignment, translation.

**Стаття надійшла до редакції 10.09.2012**