KORPUSZNYELVÉSZETI ALAPFOGALMAK

(magyar, angol, francia, német és japán nyelven, magyar meghatározásokkal)[33]

M: **Adatbázis**
A: database
F: base de données (f)
J: データベース
Strukturált / rendezett információt tartalmazó egység (pl. telefonkönyv, könyvtári katalógus), mely lehetővé teszi az adatok kívánt szempontok szerinti gyors lekérdezését.

M: **adatvezérelt nyelvtanulás**
A: data-driven language learning (DDL)
F: apprentissage de langue dirigé (induit) par les données linguistiques
N:
J: データ駆動型学習
Tim Johns, a Birminghami Egyetem tanára alkotta a kifejezést. Az autentikus nyelvi példákra épülő, gyakran konkordanciákat felhasználó nyelvtanulásra és tanításra használt kifejezés.

M: **ágbank** / szintaktikai adatbázis
A: treebank / syntactic database
F: banque d'arbres / base de données syntaxiques
N: Treebank (f) / syntaktisch analysiertes Korpus
J: 構文解析コーパス
Szintaktikailag elemzett mondatok adatbázisa.

M: **ágrajz**
A: tree diagram
F: arbre (m)
N: Baumdiagramm (n)
J: 樹形図
A szintaktikai elemzések vizuális megjelenítése, mely fára emlékeztet.

M: **általános korpusz**
A: general corpus
F: corpus de référence (m)
N: allgemeines Korpus
J: 汎用コーパス
A teljes nyelvhasználat vizsgálata céljából készült, így elsősorban lexikográfiai vizsgálatokra használják.

M: **annotáció**
A: annotation
F: annotation (f)
N: Annotation (f), Annotierung (f)
J: 1）テキスト上に付帯情報をつけること　2）情報付与
Olyan információ, mely az eredeti szövegben nem szerepel, hanem a szövegfeldolgozás során kerül a szövegbe. A szövegre vonatkozik, de a szövegtől egyértelműen megkülönböztethető. Az annotáció leggyakoribb formái a címkézés és a szintaktikai elemzés, de a nyelvtanulói korpuszok hibakódjai is idetartoznak.

M: **annotált korpusz**
A: annotated corpus
F: corpus annoté / étiqueté (m)
N: annotiertes Korpus
J: 情報付与コーパス
A szövegre és egységeire (bekezdés, szavak stb.) vonatkozó információval ellátott korpusz.

M: **átírás**
A: transcription

[33] A fogalomtár a korpusznyelvészetben használt alapfogalmakat tartalmazza. A nyelvtudomány más területeiről már ismert nyelvészeti fogalmak itt nem szerepelnek, hiszen azokat más magyar nyelvű könyvekben is meg lehet találni, pl. *Nyelvi fogalmak kisszótára* (Kugler & Tolcsvai Nagy 2000) vagy *A nyelv enciklopédiája* (Crystal 2003). A fogalomtárat folyamatosan bővítjük a www.korpusz.com lapon, így a korpusz@hotmail.com címre küldött javaslataikkal és észrevételeikkel segíthetik ennek fejlesztését és pontosítását.

F: transcription (f)
N: Transkription (f)
J: 転写
A beszélt nyelvi szövegek lejegyzése, mely több-kevesebb részletességgel történhet. A legegyszerűbb formája a puszta szavak átírása, de legtöbbször az elhangzott szavak mellett az élőbeszéd lehető legtöbb jellemzőjét igyekszik megragadni, pl. a szünetet, hangerősséget, hanglejtést, hangsúlyt stb.

M: **automatikus szintaktikai elemzés**
A: parsing
F: analyse syntaxique automatique – parsage (kanadai francia nyelvből)
N: Parsing
J: 構文解説; 構文解析
Az annotáció egy fajtája, melynek során a szöveget és a mondatokat szintaktikai egységekre bontják.

M: **automatikus szintaktikai elemző program**
A: parser
F: parseur (m)
N: Parser (m)
J: 構文解説プログラム; 構文解析ツール
A szintaktikai elemzést végző számítógépes program.

M: **beszédfelismerés**
A: voice recognition
F: reconnaissance (f) vocale
N: Spracherkennung (f)
J: 音声認識
Az a technológia, amely a hangot felismeri és automatikusan szöveggé alakítja.

M: **címke**
A: tag
F: étiquette (f)
N: Tag (igenévként: getaggtes Korpus)
J: 言語的標識
Az annotáció során a szövegegységhez kapcsolódó és azt jellemző kód.

M: **címkézés**
A: tagging
F: étiqueter
N: (Wortarten-)Tagging (n), Zuordnung (f) von Tags
J: 標識付け
Az annotáció egy fajtája, amikor a szöveg egységeit, legtöbbször a szavakat, egy vagy több, az adott egységre vonatkozó információt tartalmazó megjegyzéssel látják el. Legtöbbször a szófaji címkézést értik alatta.

M: **címkéző program**
A: tagger
F: étiqueteur (m)
N: Tagger (m), -n (plur.)
J: タグ付けプログラム
A címkézést automatikusan végző program.

M: **COCOA utalások**
A: COCOA reference
F: référence de COCOA
N:
J: COCOA 形式
Olyan hegyes zárójel-pár (<>), amely egy információkódot és annak egy valós megnyilvánulását tartalmazza. Pl. ha C=cím, akkor <C Egri csillagok>.

M: **csontváz elemzés** (szintaktikai)
A: skeleton / shallow parsing
F: parsage (m) squelettique
N: skeleton parsing, flache Analyse
J: 簡略解析
Részleges szintaktikai elemzés, amelyben egyes csoportokat további részelemekre lehetne bontani.

M: **dinamikus korpusz**
A: dynamic corpus
F: corpus dynamique
N: dynamisches Korpus
J: 動的コーパス
Folyamatosan bővülő korpusz, ahol a szövegtípusok aránya nem állandó.

M: **előfordulás**
A: occurrence
F: occurrence (f)
N: Vorkommen (n)
J: 出現
Egy adott szóalak adott korpuszban való előfordulásának a száma

M: **fordítási ekvivalencia / megfelelés**
A: translation equivalence

F: équivalence (f) de traduction
N: Entsprechung (einer Übersetzung)
J: 等価
A különböző fordításelméletek különbözőképpen határozzák meg e kulcsfogalmat. *A* nyelvről *B* nyelvre való fordításkor a fordítási megfelelőket a hétköznapi életben a kétnyelvű szótárakban keressük, de a szövegek közötti megfelelésnek a diskurzus szintjén kell létrejönni. A gépi fordítások „furcsaságai" bizonyítják ezt a legszemléletesebben. A párhuzamos korpusz esetében sem szavakat, hanem jelentésegységeket jelölnek meg fordítási megfelelőként.

M: **fordítási korpusz**
A: translation corpus
F: corpus de traduction
N: Übersetzungskorpus (n)
J: 翻訳コーパス
Kizárólag fordítás eredményeként létrejött szövegeket tartalmaz.

M: **Hapax legomenon / legomena**
A: Hapax legomenon, pl. legomena
F: hapax (m)
N: Hapaxlegomenon (n)
J: ハパックス
A korpuszban egyszer előforduló szóalak.

M: **homográf**
A: homograph
F: homographe (m)
N: Homograph (m)
J: 同綴異議語
Azonos írásképű, de különböző jelentésű szó.

M: **idióma elv**
A: idiom principle
F: principe (m) d'idiome
N:
J: 慣用原則
Sinclair nyelvelméleteben az az elv, mely szerint a jelentés szóegyüttesekből és nem különálló szavak együttes jelentéséből származik, és ezek a memóriában is így tárolódnak, vö. szabad választás elve.

M: **jelentés egyértelműsítés**
A: word sense disambiguation
F: désambiguïsation lexicale
N: Wortsinndisambiguierung (f)
J: 曖昧性除去、両義性除去
A szövegben szereplő azonos alakú, de különböző jelentésű szavak adott esetben érvényes jelentésének meghatározása. (pl. a *vár* szó főnévi vagy igei jelentése)

M: **kiegyensúlyozott korpusz**
A: balanced corpus
F: corpus équilibré
N: balanciertes Korpus
J:
Olyan korpusz, amelyben az egyes szövegtípusok és azok mennyisége többé-kevésbé hűen tükrözik a valóságban betöltött arányokat. Az általános nyelv leírásának céljából készül, így általában hatalmas mennyiségű szöveget foglal magába, melyeket elsősorban lexikográfiai célokra használnak. Lásd: általános korpusz

M: **klón**
A: clone
F: clone (m)
N: Klon (m)
J: クローン
Egy bizonyos korpusz szerkezetét követő korpusz. Legtöbbször a Brown Korpusz mintájára készült korpuszokra utal e kifejezés, de bármely más korpusz mintáját követőre is alkalmazható.

M: **kolligáció**
A: colligation
F: colligation (f)
N: Kolligation (f)
J: コリゲーション,
単語と品詞との共起パターン
Bizonyos szavak bizonyos nyelvtani szerkezetben való előfordulása, mely a valószínűségen alapuló várható véletlen együttes előfordulásnál magasabb, és sok esetben előre kitalálható, pl. *elhatározta, hogy*; vö. kollokáció

M: **kollokáció**
A: collocation
F: collocation (f)
N: Kollokation (f)
J: 連語、コロケーション,

単語の共起パターン
Bizonyos szavak együttes előfordulása, mely a valószínűségen alapuló várható véletlen együttes előfordulásnál magasabb, és sok esetben előre kitalálható, pl. *szőke haj*; vö. kolligáció

M: **konkordancia**
A: concordance
F: concordance (f) ou ligne de contexte
N: Konkordanz (e)
J: コンコーダンス
Egy adott szó vagy kifejezés szövegben szereplő összes előfordulását szövegkörnyezetében bemutató lista.

M: **konkordanciaprogram**
A: concordancer
F: logiciel de concordances (m)
N:
J: コンコーダンサー／コンコーダンス・プログラム
Konkordanciákat létrehozó program.

M: **korpusz**
A: **corpus pl. corpora**
F: corpus (m)
N: Korpus (n), Korpora
J: コーパス
Nyelvészeti vizsgálatok céljából, bizonyos szempontok alapján összeválogatott írott vagy beszélt nyelvi szövegek gyűjteménye.

M: **korpusz alapú**
A: corpus based (adj)
F: basé, -e sur le coprus
N: korpusanalytisch basiert
J: コーパスに基づく
Bizonyos problémák vizsgálatára a korpuszelemzés eredményei adnak választ, a nyelvészeti vizsgálatok a korpuszelemzésre épülnek.

M: **korpusz informált**
A: corpus informed
F: informé, - e par le corpus
N:
J:
Nem kizárólagosan korpuszelemzésre épülő kutatás. A korpuszvizsgálatok eredményeit csak megerősítésként igénylik.

M: **korpusznyelvészet**
A: corpus linguistics
F: linguistique (f) de corpus
N: Korpuslinguistik (f)
J: コーパス言語学
Azon nyelvészeti irányzat, mely a nyelv és nyelvhasználat vizsgálatát speciális módszerek és számítógépes programok segítségével korpuszra alapozva végzi.

M: **korpuszvezérelt**
A: corpus driven
F: conduit -e par le corpus
N:
J: コーパス駆動的
Az empirikus kutatásnak az a változata, amely a korpusz elemzése folyamán felfedezett szabályszerűségek alapján von le következtetéseket.

M: **KWAL formátum**
A: Key word and line (KWAL)
F: mot clé et ligne en contexte
N:
J: キーワードとコンコーダンス・ラインを中心に前後に文脈を表示する
E formátumban a keresett szót tartalmazó sor több sort kitevő szövegkörnyezetével együtt látható.

M: **KWIC formátum**
A: key word in context KWIC
F: mot clé en contexte
N:
J: キーワードを中心に前後に文脈を表示する
E formátumban a keresett szó a szövegkörnyezetével együtt látható, mely általában a számítógép monitorán egy sort tesz ki. A keresett szó általában középen helyezkedik el.

M: **lemma**
A: lemma
F: lemme
N: Lemma (n)
J: 見出し語
Az azonos szótőből származó összes (általában azonos szófajú) szóalakot átfogó kategória, pl. *ugrál, ugrik, ugrott* stb. A kutatás igényeihez igazodva különböző

szófajú alakok is tartozhatnak egy lemmába.

M: **lemmatizálás**
A: lemmatization
F: lemmatisation (f)
N: Lemmatisierung (f)
J: 見出し語化
A különböző szóalakok lemmákba való csoportosítása.

M: **lemmatizáló program**
A: lemmatizer
F: logiciel de lemmatisation
N: Lemmatiser (m)
J:
A különböző szóalakok lemmákba való csoportosítását automatikusan végző program.

M: **lexika**
A: lexis
F: lexique (f)
N: Lexika (plur.)
J: 語彙目録 語彙
Egy nyelv szókészlete, általában a nyelvtannal szembeállítva használatos.

M: **lexikon**
A: lexicon
F: léxicon (m) lexique (m)
N: Lexikon (n), Wortbestand (m)
J: 用語集
Jelentése gyakorlatilag szótár vagy szókincs, de a számítógépes programok szókincs adatbázisaira is legtöbbször ezt a kifejezést használják.

M: **mesterséges intelligencia**
A: AI artificial intelligence
F: Intelligence (f) Artificielle
N: künstliche Intelligenz
J: 人工知能、AI
Az emberi agy működését vizsgálja azzal a céllal, hogy számítógépek segítségével szimulálja azt.

M: **MI együttható**
A: MI score (mutual information)
F: score (m) d'information mutuelle
N:
J:（特定の2語間の連想関係の強さを計る尺度）
Két szó **tényleges** együttes előfordulását a valószínűségi számítások alapján **„várható"** előfordulással való összehasonlítás eredménye (informatikai elméletből származik). A kollokációk vizsgálatánál használják.

M: **mintavétel**
A: sampling
F: échantillonnage (m)
N:
J: 標本抽出 サンプリング
Szövegminták kiválasztása adott populációra (szövegtípusra, regiszterre) vonatkozó vizsgálatok végzése céljából.

M: **monitor korpusz**
A: monitor corpus
F: corpus de suivi (de baromètre)
N: Monitorkorpus (n)
J: モニター•コーパス
Olyan korpusz, amely lehetővé teszi a nyelv rövidtávú változásának elemzését a szövegek korpuszon belüli elkülönítésével. Rendszeresen fejlesztik a korpuszt, de az arányokat mindig megtartják.

M: **n-gram**
A: n-gram
F: n-gramme (f)
N:
J: (文字列の連鎖を集計する機能)
Az n-gramok a szövegben szereplő több egységből álló szerkezetek/szövegsablonok korpuszvezérelt elemzését segítik elő. Az *n* helyére kerülő szám határozza meg, hogy a szövegben szereplő szavakat hányas egységekbe csoportosítjuk. Pl. *A macska az asztalon ül.* 3-gram esetében: *a macska az, macska az asztalon,* és *az asztalon ül* egységekre bonthatók.

M: **nyelvtanulói korpusz**
A: learner corpus
F: corpus d'étudiants de langue
N:
J: 学習者コーパス
Nem anyanyelvi beszélők nyelvi megnyilvánulásaiból összeállított korpusz.

M: **összehasonlítható korpusz**
A: comparable corpus

F: corpus comparable
N:
J: 比較コーパス・コンパラブルコーパス（多言語間または複数言語変種間の比較ができるように、同じコーパスデザインで編纂されたコーパス）
Két vagy több olyan korpusz, amelynek szerkezete és mérete hasonló (pl. angol és magyar nyelvű üzleti levelezés).

M: **párhuzamos korpusz**
A: parallel corpus
F: corpus parallèle
N: Parallelkorpus (n)
J: パラレル・コーパス（2 カ国語以上による同じ内容のコーパス）
Olyan korpusz, amely több mint egy nyelven tartalmazza ugyanazt a szöveget vagy szövegeket.

M: **pedagógiai korpusz**
A: pedagogic corpus
F: corpus pédagogique
N: pädagogisches Korpus
J: 教育型コーパス
Dave Willis szóhasználatában azon szövegek összessége, amellyel a nyelvtanulók egy kurzus alatt találkoznak.

M: **példány**
A: token
F: occurrence (f)
N: Token (n)
J: トークン(総語数)
Egy szövegben akár többször is előforduló bármely szó; vö. szövegszó.

M: **probabilisztikus módszerek**
A: probabilistic methods
F: méthodes probabilistes (f)
N: Wahrscheinlichkeitsmethoden
J: 確率論的手法
A statisztikai valószínűségen alapuló módszerek.

M: **reguláris kifejezés (regexp)**
A: regular expression
F: expression (f) régulière
N: regulärer Ausdruck (m)
J: 正規表現
A programozásban használt olyan kifejezés, amelyet főleg szűrőknél, minták feldolgozására és keresésére használnak.

M: **reprezentatív**
A: representative
F: représentatif, -ive
N: repräsentativ
J: 代表的 典型的な
Olyan minta, amely a populációra jellemző jegyek összességét a lehető legnagyobb mértékben megközelíti.

M: **SGML szabványos, általános leíró nyelv**
A: SGML Standard Generalized Markup Language
F: language standard de balisage SGML
N:
J: 形式
Szöveges állományok belső szerkezetének (fejezetek, bekezdések, lábjegyzetek stb.) jelölésére használható szabvány.

M: **statikus korpusz**
A: static corpus
F: corpus statique
N: statisches Korpus
J:
Olyan korpusz, amelynek tartalma nem változik.

M: **statisztikai jelentősség**
A: significance
F: significance (f)
N: Signifikanz (f)
J: 有意水準 (level)
Nem a véletlenen múló statisztikai eredmény, mely alapján következtetések vonhatók le.

M: **szabad választás elve**
A: open choice principle
F: principe (m) de choix
N:
J:
Sinclair szóhasználatában az idióma elvvel ellentétes elv, mely szerint a jelentés az egyes szavak jelentésének összességéből jön létre, így minden szó után szabadon választható meg a következő, vö. idióma elv.

M: **számítógépes nyelvészet**
A: computational linguistics

F: linguistique informatisée?
N: Komputerlinguistik (f)
J: コンピュータ的言語学
A nyelv vizsgálatához számítástechnikai elveket és módszereket használó tudományterület.

M: **szemantikai prozódia**
A: semantic prosody
F: prosodie sémantique
N: semantische Prosodie
J: 意味的プロソディ
Bizonyos szavak csak bizonyos jelentéstartalmú szavakkal vagy nyelvi szerkezetekkel együtt fordulnak elő. Pl. az ***okoz*** általában negatív dolgokkal: *bánat, baleset, kár* stb.

M: **szóalak/típus**
A: type
F: type (m)
N: Typ (m)
J: タイプ(異なり語数)
A szövegben előforduló különböző írásképű szó (pl. bot, botot).

M: **szókincstár (thesaurus)**
A: thesaurus
F: thésaurus (m)
N: Thesaurus (m)
J: 類語辞典, シソーラス、語彙分類集
Olyan szótár, amelyben a szavakat jelentésük alapján csoportosítják, nem pedig ábécérendben.

M: **szöveggyűjtemény /szövegarchívum**
A: text collection/text archive
F: collection de textes/archive de textes
N: Textarchiv (n)
J: テキスト集合体
Szövegek esetleges gyűjteménye, tárháza; vö. korpusz.

M: **szövegkörnyezet**
A: context
F: contexte (m)
N: Kontext (m)
J: 文脈,前後
A korpusznyelvészetben használt meghatározás szerint egy szót vagy kifejezést közvetlen megelőző és követő szövegrészlet. Ennek segítségével lehet az adott szó vagy kifejezés jelentését egyértelműsíteni.

M: **szövegszinkronizálás**
A: alignment
F: alignement (m)
N: Alignierung (f)
J:
Két szöveg/korpusz egymáshoz való igazítása, melynek során az összetartozó elemeket megjelölik. Így az egyikben történő kereséskor a másikban hozzárendelt adatok is megjelennek. Fordítási vagy párhuzamos korpuszok esetében igen gyakori.

M: **szövegszó**
A: running words
F: mot (graphique)
N: laufende Wörter/Wortformen
J:
Olyan betűcsoport, amelyet mindkét oldalon szóköz választ el. Esetenként a jobb oldali szóközt írásjel előzi meg.

M: **T-együttható, Tí-szkór**
A: T-score
F: score T de cooccurrence
N:
J: T-score (特定の2語間に j 何らかの連想関係があることを主張することができる確信度を計る尺度)
Két elem közötti összefüggésre vonatkozó, statisztikai együttható, a kollokációk vizsgálatánál használják.

M: **TEI (szövegkódolási ajánlás)**
A: (TEI) Text Encoding Initiative
F: TEI (f) Initiative de documentation de textes
N:
J: テキスト電子化の規格化
Olyan nemzetközi és interdiszciplináris szabvány, amely a szövekódolást igyekszik egységessé tenni. A pontosságra és egyszerűségre törekednek. A szabvány fejlesztésére 1987-ben konzorciumot hoztak létre.

M: **teljes szintaktikai elemzés**
A: full parsing
F: analyse (f) complète
N:
J: 詳細解析

A mondat minden egységének legkisebb szintaktikai egységre való lebontása.

M: **Természetes nyelvfeldolgozás**
A: NLP Natural Language Processing
F: traitement automatique du language naturel (m) (TALN)
N:
J: 自然言語処理
A nem formális, azaz emberi nyelvek számítógépes feldolgozása.

M: **többnyelvű korpusz**
A: multilingual corpus
F: corpus multillingue
N: mehrsprachiges Korpus
J: 多言語コーパス
Több nyelven tartalmaz szövegeket. Ezek típusuk szerint különbözők lehetnek: fordítási, összehasonlítható vagy párhuzamos korpusz.

M: **többváltozós statisztikai elemzés**
A: multivariate statistics
F: analyse (f) statistique multivariée
N: multivariate Statistik
J: 多変量統計
Olyan statisztikai elemzések, amelyek egyszerre több változó közötti kapcsolatokat vizsgálnak.

M: **történeti/ diakronikus korpusz**
A: historical/diachronic corpus
F: corpus diachronique
N: historisches/diakronisches Korpus
J: 通時コーパス
A nyelv történeti változásának tanulmányozása céljából nem kortárs szövegeket tartalmazó korpusz.

M: **tudásbázis**
A: knowledge base
F: base de connaissance (f)
N: Knowledge-Base (f)
J:
Egy mesterséges intelligencia program műveletek elvégzéséhez szükséges szabályainak forrása, melyek formális nyelven íródnak.

M: **tudatosság javítása / tudatosságot javító**
A: awareness/consciousness raising
F: élévation (f) du niveau de conscience
N: Bewusstseinerhebung (f)
J: 気づき／意識化
A nyelvtanítás azon elve, mely szerint a nyelvtan explicit tanítása helyett a nyelvtanuló nyelvi/nyelvtani tudatát olyan feladatokkal emeli, mely a diákok részéről aktív megfigyelést és következtetések levonását igényli.

M: **vizsgált csomópont/adat (nód)**
A: node
F: noeud (m)
N:
J: 中心点
A konkordanciákban és az elemzésekben a vizsgált adat helyett gyakran megjelenő kifejezés.

M: **Z-együttható, Zí-szkór**
A: Z-score
F: score Z de cooccurrence
N:
J: Z 値
Két elem közötti összefüggésre vonatkozó, statisztikai együttható, a kollokációk vizsgálatánál használják.