Váradi Tamás, a Nyelvtudományi Kutatóközpont tudományos főmunkatársa 1976-ban szerzett angol–spanyol szakos nyelvtanári diplomát, valamint általános és alkalmazott nyelvészet szakos bölcsész oklevelet az Eötvös Loránd Tudományegyetemen. 1973-tól 1983-ig tanársegédként, majd adjunktusként oktatott a Külkereskedelmi Főiskolán. 1987-től az MTA Nyelvtudományi Intézetének tudományos munkatársa. 1990 és 1991 között a Lancesteri Egyetem vendégkutatója volt, majd 1991–1995 között a Londoni Egyetem Szláv és Kelet-Európai Tanulmányok Intézetében dolgozott nyelvi lektorként. 1997-ben PhD-fokozatot szerzett angol nyelvészetből, s ugyanez évtől látja el az osztályvezetői feladatokat az MTA Nyelvtudományi Intézetének nyelvtechnológiai osztályán.

Váradi Tamás 2021 márciusában töltötte be 70. életévét. Ezzel az ünnepi kötettel szeretnénk tisztelegni munkássága előtt.

A KORPUSZNYELVÉSZETTŐL A NEURÁLIS HÁLÓKIG

A KORPUSZNYELVÉSZETTŐL A NEURÁLIS HÁLÓKIG

KÖSZÖNTŐ KÖTET VÁRADI TAMÁS 70. SZÜLETÉSNAPJÁRA

Szerkesztette:
Dodé Réka
Ludányi Zsófia

Nyelvtudományi Kutatóközpont
Budapest, 2021

A korpusznyelvészettől a neurális hálókig

Köszöntő kötet
Váradi Tamás 70. születésnapjára

A korpusznyelvészettől a neurális hálókig

Köszöntő kötet
Váradi Tamás 70. születésnapjára

Szerkesztette:
Dodé Réka
Ludányi Zsófia

Nyelvtudományi Kutatóközpont
Budapest, 2021

Az angol nyelvű szövegeket lektorálta:
Tóth Dániel Miklós

ISBN: 978-963-9074-90-3

A kiadásért felelős a Nyelvtudományi Intézet igazgatója
Megjelent a Nyelvtudományi Intézet gondozásában

Nyomdai előkészítés: Ligeti-Nagy Noémi

A borító Váradi Tamás fotójának felhasználásával készült.

Megjelent: 2021-ben

Készült az EFO Kiadó és Nyomda Kft. nyomdájában, Százhalombattán
Felelős vezető: Fonyódi Ottó

Tartalomjegyzék

TABULA GRATULATORIA

Alexin Zoltán
Bakró-Nagy Marianne
Bartha Csilla
Dodé Réka
Dömötör Adrienne
Fenyvesi Anna
Fóris Ágota
Gerstner Károly
Gósy Mária
Gráczi Tekla Etelka
Gyarmathy Dorottya
Győrffy András
Héja Enikő
Heltainé Nagy Erzsébet
Horváth Viktória
Hunyadi László
Huszár Anna
Jelencsik-Mátyus Kinga
Kalivoda Ágnes
Kardos Tamás
Kenesei István
Kirchmeier, Sabine
Koeva, Svetla
Krepsz Valéria
Lendvai Piroska
Ligeti-Nagy Noémi
Lipp Veronika
Ludányi Zsófia
Makrai Márton
Markó Alexandra
Mittelholcz Iván
Navracsics Judit
Oravecz Csaba
Oszkó Beatrix
Prószéky Gábor
Raátz Judit
Sass Bálint
Sherwood, Peter
Simon Eszter
Simon László
Sipos Mária
Stickel, Gerhard
Surányi Balázs
Szécsényi Tibor
Szekrényes István
Tadić, Marko
Takács Dávid
Tufiş, Dan
Vadász Noémi
Várnai Zsuzsa
Vincze Veronika

Váradi Tamás, a gentleman

Kenesei István[1]

[1] Nyelvtudományi Intézet
kenesei.istvan@nytud.hu

Ha Váradi Tamást röviden kellene jellemezni, akkor azt mondanám, az előtte lebegő minta egy angol lord, de legalábbis egy brit gentleman, annak elegáns tweedöltönyökből álló ruhatárával, szókincsével, valamint számos hobbijával, amelyek között Tamás esetében a lovaglás helyett a biciklizés és a – pár lóerővel modernebb – motorozás mellett a nyelvvel való bíbelődés is előkelő helyen állna, valahogy úgy, ahogy a példaképe számára mondjuk a tájképfestés. A rókavadászatról bizonyára szívesen lemondana, de a kandalló mellett, hű kutyája fejét simogatva és whiskyjét szopogatva olvasná a Timesnak az inas által frissen vasalt lapjait. Ne feledjük, hogy ez az anglofil vonás nem idegen a magyar értelmiségiektől, most csak a nyelvészettől is megérintett kortársainkra gondolva, Országh Lászlótól kezdve egykori tanárunkon, András T. Lászlón át egészen közös barátunkig, Nádasdy Ádámig.

A mi szerencsénk az volt, hogy Tamásnak a nyelvészet nem lehetett a hobbija, mert szemben Albion úriembereivel neki megélhetés után kellett néznie, és ha már kerékpár- vagy motorversenyző nem lehetett, ami amúgy sem fért volna össze türelemre hajló személyiségével, akkor a nyelvtudománnyal kötelezte el magát. Ráadásul még idejekorán, szinte egyetemista korában, amikor az 1970-es évek elején megalakult angol–magyar kontrasztív kutatáshoz csatlakozott, amelyet William J. (Bill) Nemser (1923–2015) vezetett közösen Dezső Lászlóval (1928–2016), és ennek köszönhetően került ki az Egyesült Államokba is. Később nagy szerencse érte, hogy a (legalábbis számunkra) híres, mert egyetlen egyesült királyságbeli magyar tanszékre került a University of London School of Slavonic and Eastern European Studies patinás épületébe a Russell téren, közel a British Museumhoz. Sajnos ez a kicsiny, de fontos tanszék, mondhatjuk, a rendszerváltás áldozata lett, és mivel – más kelet-európai kultúrák támogatóival szemben – nem akadt olyan önzetlen milliomos, beleértve a magyar államot is, aki alapítványt tett volna, hogy tovább mű-

DOI: https://doi.org/10.18135/VT70.1

ködhessen, nemcsak elköltözött a Faber & Faber Kiadó T. S. Eliot emléktáblájával is megjelölt egykori otthona mellől, de lassan teljesen elsorvadt.

Tamás ekkor már az MTA Nyelvtudományi Intézete kötelékében dolgozott, ahol – Herman József igazgató (1924–2005) terveit megvalósítandó – a nyelvtörténeti kutatások megalapozása céljából alakult szociolingvisztikai csoport tagja lett. A korpuszépítés feladata vezette a számítógépes nyelvészet felé, s előbb a korpusznyelvészeti csoport vezetője lett, majd egyre jobban beletanulva a nyelvtechnológiába, nemcsak azt ismerte fel, hogy a saját jövőjét itt találja meg, hanem azt is, hogy a következő nyelvésznemzedékeknek micsoda tejjel-mézzel folyó Kánaánja, azaz kimeríthetetlen kutatási, innovációs és fejlesztési területe jöhet létre itt. Arra is hamar ráébredt, hogy ha idejében kialakítja nemzetközi kapcsolatait, akkor a hazai nyelvészeket a tudományág felfutásának a kezdetén összeköthetti a szakma élvonalával. Az már a mi szegény országunknak és alulfizetett nyelvészeinek a sorsa, hogy az így menedzselt kiválóságokat egyre-másra találják meg azok a nemzetközi cégek, amelyek magas béreket és kiváló munkafeltételeket, valamint problémagazdag feladatokat ígérve rendre elcsábítják az általa felfedezett tehetségeket. Magam is tanúja voltam, ahogy sorban vonultak az EU fordítói központjába vagy külföldi egyetemekre, akikre boldogan számítottunk volna magunk is. De ez Tamásnak sohasem szegte a kedvét: felállt, megrázta magát és próbálta kinevelni a következő csapatot. Azután, ahogy az Intézet belső szerkezete átalakult, egyre jobban kellett (és lehetett!) támaszkodni rá.

Amikor 2004-ben az ESFRI (European Strategy Forum for Research Infrastructures) magyar delegátusa lettem, Tamás aktív közreműködésére is támaszkodhattam, hogy az általa társmenedzselt TELRI hálózat addigi eredményei alapján pályázzon az EU által támogatandó kutatási infrastruktúra státuszára. Az egyeztetések után, melyekben az anyagilag jóval fölöttünk lévő súlycsoportba tartozó Utrechti Egyetem és nijmegeni Max Planck Intézet vett részt, Tamás hathatós részvételével alakult meg a máig egyetlen és ma már megkerülhetetlen CLARIN (Common Language Resources and Technology Infrastructures) európai kutatási infrastruktúra, amelyben jelenleg 23 európai ország képviselteti magát. Emlékezetes, ahogy a megalakulás nagy sikere után évek hosszú küzdelme várt ránk, hogy a változó felállású és vezetésű kutatási hivatal végre befogadja ezt az egyetlen humán tudományi kutatási infrastruktúrát és kifizesse a más tudományterületekéhez képest minimális évi tagdíjat, aminek köszönhetően pár éve végre teljes jogú tagokká válhattunk.

Ezzel a háttérrel közvetít Tamás mind a mai napig az ország vezető nyelvtechnológiai szereplői között, hozott létre egy egyedülálló nyelvtechnológiai platformot és pályázik kitartóan hazai és EU forrásokra.

Tamásban van egy további tehetség is: a kimeríthetetlen szervezői tettvágy és invenció. Soha nem ad vissza megbízást, soha nem tagad meg kérést, soha nem hárít el egy „muszáj-feladatot". Nem véletlen, hogy ideális igazgatóhelyettes lett belőle: tárgyalóképes és megbízható képviselője kis és nagy ügyeknek egyaránt. Alkalmazott nyelvészeti szemináriumoktól egészen a monumentális 2011. évi budapesti META-FORUM-ig bármit professzionális szinten szervez meg. Találékonyságára pedig legyen példa a *helyesiras.mta.hu* lapon látható nagy látogatottságú automatikus helyesírás-ellenőrző program, amelyet az Akadémia egyik korábbi elnöke maga avatott fel.

És ő volt az is, akit 2004-ben megkértem, hogy kémlelje ki, milyen intézmény az EFNIL, vagyis az Európai Nyelvi Intézetek Szövetsége, ahova oly elszántan invitáltak minket, de amitől kicsit tartottunk, mert tudományos kutatóhely lévén nem akartunk, mondjuk, egy nemzetközi nyelvvédő szervezetbe belépni. Tamás egy párizsi kiküldetés során fényesen teljesítette a feladatot és ahogy az NYTI egyre jelentékenyebb tagjává vált a szervezetnek, Tamás szerepe is fokozatosan megnőtt: előbb 2007-től neki köszönhetően a NYTI lett a honlapgazda, majd amikor 2009-től az EFNIL titkárságának is az otthona lettünk, őt választották a szervezet főtitkárává, s ezzel mintegy a második élete is elkezdődött, amibe teljes elánnal vetette bele magát – az EFNIL szerencséjére mind a mai napig. Meggyőződésem, hogy nem kis részben neki, pontosabban felkészült külügyéreket is megszégyenítő diplomáciai képességeinek tulajdonítható, hogy az EFNIL ennyire zökkenőmentesen halad tovább a még szélesebb európai együttműködés irányában. Szerencsére Tamás szenvedélyes utazó is: amíg másnak a háta borsódzik az olyan átszállásoktól, akár éjszaka is, amelyek miatt órákat kell a repülőterek korántsem barátságos termeiben dekkolni, neki ez csak múló kellemetlenség, hiszen tűzbiztos e-felszereléssel utazik, amivel bárhol és bármennyi ideig el tudja foglalni magát.

Mint az Intézet pályázati és egyéb bevételekkel talán legjobban ellátott részlegének vezetője mindig készségesen kisegíti a nagy egészet, ha szükség van rá. Évekig neki volt köszönhető, hogy a szűkösen csordogáló akadémiai támogatású konferencia-részvételeket kiegészíthettük intézeten belüli utazási pályázatokkal. De támaszkodhattunk rá minden esetben, amikor újra és újra el kellett készíteni a NYTI pályázatait azért,

hogy egyáltalán megkapjuk a fenntartásunkra szolgáló összegeket, hogy hozzájuthassunk a működésünkhöz szükséges infrastruktúrákhoz, hogy kidolgozzuk fejlesztési terveinket – és ezek rendszeresen a nyár közepére voltak időzítve, mintegy gondoskodva róla, hogy ne kényelmesedjünk el a nagy melegben.

Tamás kiváló kapcsolatépítő és kapcsolattartó e legkülönfélébb intézményekkel és szervekkel. Máig emlékszem a magyarországi cigányság által beszélt nyelvek elektronikus feldolgozásának témájára, amelyen kormányzati szereplőktől az IBM-ig vettek részt fontos személyek. Nemegyszer ez az aktivitása vissza is ütött, mert nem mindig tudott ellentmondani a felkéréseknek, amelyekbe az NYTI-t is bevonta, vagy maga talált ki olyan lobbizási lehetőségeket, melyek aztán nem hozták a várt eredményt. Máig emlékszem kudarcosnak is mondható látogatásunkra a nem túl nagy népszerűségnek örvendő oktatási államtitkárnál, aki fegyelmezetten végighallgatott minket, majd semmit sem ígérve, netán alig valamit felfogva a mondandókból vett búcsút tőlünk. És arra a nemzetgazdasági minisztériumi államtitkárra is, aki olyan megfigyelésekkel látott el minket köszöntőjében, hogy: „angol lingvisztikusok szerint pedig: ha minden nyelv egy drog lenne, akkor a magyar a varázsgomba” vagy hogy a miénk „a világ harmadik legnehezebben megtanulható nyelve”, valamint az olasz és a görög után „a harmadik legdallamosabb nyelv a világon”.

Nem véletlen, hogy az Intézetet ért minden csapás és nehézség ellenére Tamás meg tudta őrizni a humorérzékét, amire attól tartok, a jövőben is nagy szüksége lesz.

Tamás Váradi at seventy: the London years

Peter Sherwood[1]

[1] László Birinyi, Sr., Distinguished Professor (Emeritus) of Hungarian Language and Culture, University of North Carolina at Chapel Hill,
School of Slavonic and East European Studies
magyarize@gmail.com

It is difficult enough, in the middle of the Covid-19 pandemic, to recall what life was like even a year ago, let alone a generation ago, in the final decade of the twentieth century. This is especially true of life in tertiary education in the UK, where changes have been extraordinarily rapid and what detailed documentary evidence remains is unusually sparse. I hope, therefore, that I will be forgiven for offering more background than might be expected in this celebration of my good friend Tamás Váradi's years of teaching in London.

The bilateral cultural agreement between Great Britain and Hungary in force between 1964 and 1998 provided for a teacher of Hungarian to be placed at the only university in the country that at the time offered a full degree in Hungarian, namely the School of Slavonic and East European Studies (SSEES), then an independent constituent institute of the academic umbrella known as the University of London and, since 1999, part of University College London. The first to hold the post of *lektor* was the English literature scholar Péter Egri; by my reckoning Tamás was the ninth such native speaker dispatched to help SSEES students primarily with their spoken Hungarian language skills, arriving in 1991 and staying until 1995. Just to clear up a common misconception: SSEES never had a Hungarian department *(magyar tanszék)* as such, only teachers of language and literature – during Tamás's tenure these were myself and Daniel Abondolo –, with Hungarian history being the responsibility of László Péter in the department of history, and aspects of Hungarian sociopolitics being taught by George Schöpflin, who held a joint post at SSEES and the London School of Economics. Tamás compiled a useful factual outline of the Hungarian teaching then offered by SSEES (Váradi, 1993). Although this brief article was preceded in the same volume by my historical piece contextualising Hungarian studies in London (Sherwood, 1993), this latter study has now been superseded

DOI: https://doi.org/10.18135/VT70.2

by an article I co-authored with my successor at SSEES (Sherwood and Tarsoly, 2008).

In the years immediately after the fall of communism in 1989, Hungary enjoyed what proved to be an all-too-brief surge of interest in the West and this manifested itself at the level of tertiary education, too: my files show that in 1991-2, Tamás's first academic year at SSEES, there was a record number of 21 students of Hungarian, at various levels and taking various degrees, and the institution was honoured with a visit by the first president of post-communist Hungary, Árpád Göncz, in November 1991. As a matter of fact, the director of SSEES, the scholar of Finnish Michael Branch, sent a memorandum to the Hungarian Foreign Ministry in 1993, proposing the establishment of a Hungarian Centre at the School, though in the end nothing came of the idea: by that time the window of Western interest in the former Soviet satellites was already beginning to close. Meanwhile, despite a considerable teaching load, in 1992 Tamás helped to establish and run a Hungarian Seminar for the relatively large numbers interested in matters Hungarian, and he was also enterprising and energetic enough to launch a Hungarian film club at SSEES in 1994.

It was always a pleasure to work with Tamás, whose rapport with the students was legendary and whose professionalism as a teacher had been honed by an apprenticeship in a Hungarian secondary school – unlike some of the *lektors* assigned to SSEES during the communist era, who though in the best cases outstanding university teachers of English, were – in many of the worst – lightweight journalists with no teaching qualifications, or undistinguished university folk who sent secret reports to the Hungarian authorities about SSEES teachers and the students studying Hungarian at SSEES (I have seen them).

Perhaps the most significant academic activity of Tamás's in London – in addition to the perfecting of his superb command of English – was his invaluable assistance with a one-time pet project of mine, the London Learner's Dictionary of Hungarian. It should be remembered that we are speaking of the years that were still largely B.C., abbreviating in this case 'Before Computers'. My idea was based on the hardly surprising recognition that the bilingual dictionaries of English and Hungarian were necessarily designed to help the very large numbers of Hungarians learning English and much less the very limited number of English-speakers likely to be interested in learning Hungarian. Unsurprisingly, an entirely

different approach is needed for the latter group, and Tamás's burgeoning computer linguistic skills were essential in formulating this, as outlined in our joint article (Sherwood and Váradi, 1993). Although I published a further article (Sherwood, 1997), on the treatment of co-verbed verbs (more traditionally 'verbs with verbal prefixes') in the proposed dictionary, it is, of course, no fault of Tamás's that, after his departure, the closing of the above-mentioned window, and not least the relentless expansion of online dictionaries based on increasingly massive computer corpora – the exploitation of which, as we know, went on to become one of Tamás's main fields of expertise – this project was one of those to fall by the wayside.

It gives me great pleasure to have this opportunity to place on record Tamás's contributions to Hungarian studies in London and, in congratulating him on his seventieth birthday, to wish him many more years of active and pleasurable research in the areas of linguistics to which he has already contributed so much.

References

Sherwood, P.: Alkalmazott nyelvészet a tanulói szótár felépítésében (Igekötős igék a londoni magyar tanulói szótárban). Hungarológia 9, 154–160 (1997)

Sherwood, P., Tarsoly, E.: A múlt mint előjáték? A hungarológiai stúdiumok hetven esztendeje Londonban. THL2: Journal of Teaching Hungarian as a Second Language/A magyar nyelv és kultúra tanításának szakfolyóirata 1–2, 5–17 (2008)

Sherwood, P., Váradi, T.: A londoni magyar tanulói szótárról. Hungarológia 4, 101–113 (1993)

Váradi, T.: Tájékoztató a Londoni Egyetem Szláv és Kelet-Európai Tanulmányok Intézetében (SSEES) folyó magyar szakos oktatásról. Hungarológia (a Nemzetközi Hungarológiai Központ kiadványa) 1, 154–157 (1993)

Tamás Váradi and EFNIL

Réka Dodé[1], Gerhard Stickel[2]

[1] Nyelvtudományi Intézet
dode.reka@nytud.hu
[2] Leibniz-Institut für Deutsche Sprache,
European Federation of National Institutions for Language
g.stickel@t-online.de

1. European Federation of National Institutions for Language

EFNIL, the European Federation of National Institutions for Language, has been an important part of Tamás Váradi's professional life for 17 years and still is. Here is a short history of EFNIL with a focus on Tamás' merits as administrator and impulse-giver.

EFNIL was founded at a conference in Stockholm in 2003 after three years of preparation at conferences in Mannheim, Florence and Brussels and several meetings of a steering committee that had been elected in Brussels. The idea was to establish a network between the central national language institutions of all member states of the European Union.

At the Stockholm conference in 2003, representatives of language academies and other central language institutions from 13 European countries decided to found the new organisation: Belgium, Denmark, Finland, France, Germany, Greece, Great Britain, Italy, Luxembourg, Netherlands, Portugal, Spain and Sweden. Representatives of language organisations from Austria, Iceland and Norway were present as observers. The founding members agreed on a constitution and the somewhat clumsy name *European Federation of National Institutions for Language* (abbreviated *EFNIL*) for the newly established organisation. Since then, the objectives of EFNIL have been:

- To promote European linguistic diversity as a means of preserving and extending the richness of European culture and developing a sense of shared European identity
- To support the European national languages as the best guarantors of linguistic opportunity within their respective member states.

DOI: https://doi.org/10.18135/VT70.3

- To support the European language organisations in their roles as centres of excellence for linguistic analysis and description and as advisory bodies on language policy to relevant political institutions.
- To facilitate the exchange of information and the development and promotion of joint European linguistic research projects between language institutions.
- To encourage, within each member state:
    - the teaching of the national language or languages at all educational levels in schools, in order to promote the written and oral competence that is necessary to enable people to play a full role in society;
    - the teaching of foreign languages within the educational system (in accordance with common European performance standards) from the earliest possible age;
    - opportunities for non-native speakers (both children and adults) to learn the national language of the state in which they reside, and also to maintain competence in their native tongue;
    - exchange opportunities for students and teachers within the European Union.[1]

The founders of EFNIL all came from member states of the 'old' European Union, that is before its various enlargements from 2004 until 2013. Only a few years after the various enlargement steps, central language institutions from the new member states of the EU also joined EFNIL. As the constitution of EFNIL allows institutions from other European countries to become associate members, language institutions from Iceland, Norway, Switzerland and Serbia also joined the federation. Over the years, a few institutes had to interrupt their membership due to financial problems. The present Executive Committee and the president of EFNIL are trying to win them back.

Tamás Varadi's involvement in EFNIL began back in 2004. Tamás and Gerhard Stickel met on 7th November 2004 in Paris. The Délégation générale à la langue française et aux langues de France, a founding member of EFNIL, was the hosting organisation. The French colleagues had

[1] See the website of EFNIL at http://efnil.org/documents/principles

chosen the castle of Sèvres as the site of EFNIL's second annual conference. Gerhard reports that Tamás and he had both arrived early for the conference. So they took a walk discussing various linguistic topics in connection with EFNIL. Tamás was highly interested in EFNIL as a European language organisation and its aims. Gerhard explained to him that EFNIL itself was not a research project but an organisational initiative to maintain the various official standard languages of the European states and to promote and develop the European linguistic diversity. EFNIL, however, could organise and support special projects that would serve its general aims.

At the Paris conference, the members of EFNIL and Tamás, as one of the observers, discussed both the general linguistic situation in several states of the European Union and the problems and solutions concerning legal terminologies in different European languages. Some of the papers read mentioned the computational means of collecting, comparing and adapting the different national terminologies, an aspect Tamás has always been especially interested in.

In the course of 2005, Istvan Kenesei, the head of the Research Institute for Linguistics of the Hungarian Academy of Sciences (Magyar Tudományos Akadémia Nyelvtudományi Intézet) and Tamás, as his deputy, decided to apply for membership of their institute in EFNIL. Both subsequently attended the annual conference 2005 in Brussels where the application was accepted by the general assembly of EFNIL with applause from the other delegates.

Like the other EU states, Hungary could now be represented by two delegates, one of these being Istvan Kenesei until 2017, the other one a number of times (Madrid 2006, Riga 2007) an official from the Hungarian Ministry of Culture and Education. Tamás has been a delegate for his institute at the other EFNIL conferences.

Since 2004, Tamás has attended most of the many annual conferences of EFNIL – the conferences in Dublin, Thessaloniki, London, Budapest, Vilnius, Florence, Helsinki, Warsaw, Mannheim, Amsterdam, and Tallinn (2019). In 2010 he was elected member of EFNIL's executive committee (sometimes called the board) and appointed General Secretary of EFNIL. The same year, the Secretariat of EFNIL was moved from The Hague to Budapest. As General Secretary, Tamás, with the support of an assistant and in close contact with the president of EFNIL, has since been responsible for the various tasks of the Secretariat, including:

- preparation of the annual conferences in cooperation with the local organisers
- finances of EFFNIL including the annual accounts
- preparation of meetings of the Executive Committee
- preparation of annual reports on EFNIL's activities and plans for further activities
- minutes of the various conferences and meetings
- website of EFNIL and its various parts including a news section
- contacts with and between the member institutions
- archive of documents pertaining to EFNIL

As Tamás proved to be the 'backbone' of EFNIL's organisation, he was re-elected General Secretary three times (2012, 2015, 2018) and has held this office now (2021) for ten years.

Beside the exchange of information between its member institutions and contacts with national and international political institutions, the activities of EFNIL have been highlighted by the annual conferences on topics related to the policies and aims of its members and EFNIL in general, and by several projects also related to EFNIL's principles. Tamás has been involved in the thematic orientation of several conferences as well as in two major projects.

Information technology, especially computational linguistics, has been a major field of Tamás' research work. He was, therefore, the highly competent co-editor of two of EFNIL's annual conference publications: the 'yearbooks' 2010 (Thessaloniki) and 2012 (Budapest) on the general topics:

- Language, Languages and New Technologies. Thessaloniki 2010 (Stickel and Váradi, 2011).
- Lexical Challenges in a Multilingual Europe. Budapest 2013 (Stickel and Váradi, 2013).

Beside his tasks as General Secretary, he was and still is especially involved in two international projects of EFNIL: EFNILEX and ELM. A short description of these projects follows.

2. EFNILEX Automatically Generated Online Dictionaries

Lexicographical resources (dictionaries, mono- and multilingual corpora) are among the means of overcoming the barriers between the various official languages and thus maintaining European linguistic diversity. The 2012 annual conference in Budapest that Tamás and his local colleagues organized was, therefore, devoted to these aims. The conference was concluded with a "Budapest Resolution". One of the core articles of this resolution reads:

> "The development of a modern high-quality lexicographical infrastructure, consisting of good (parallel) corpora, mono-, bi-, and multilingual dictionaries and lexical databases, thesauruses, terminology databases, wordnets and comparable lexical tools and instruments is a *conditio sine qua non* for the learning and use of (foreign) languages by European citizens, for translation and interpretation and for the multilingual processing of texts by technical systems and devices" (Stickel and Váradi, 2013: 212).

A project group headed by Tamás and Johan Van Hoorde (Dutch Language Union) was subsequently formed for the development of a (semi-)automatic method for the preparation of bi- and multilingual dictionaries at low cost. The method is based on parallel corpora and natural language processing tools. The output of the pilot project is a lexicographical tool that can be a substantial aid in dictionary writing. It is a proto-dictionary generated automatically based on real language data and contains all lexicographical information needed for the actual lexicographers' work.[2]

3. The European Language Monitor: ELM

In order to create an instrument to provide an empirical basis for national and European language policies, EFNIL has been designing and constructing a European Language Monitor (ELM). The motive for this project is that EFNIL realized, in the course of its activities, the lack of a satisfactory empirical basis for national and European language policy concepts and measures. The data available on the present linguistic situation of the various countries are rather heterogeneous, incomplete and, in part, outdated. The valuable results of some national projects and of

[2] See: http://efnilex.efnil.org

European surveys such as Eurobarometer and Eurydice are only a partial remedy because they are limited to foreign language learning and foreign language competence. Politicians at national and European level, language planners, educationalists, linguists, and the general public are obviously in need of a reliable and up-to-date linguistic picture of all the member states, that is, of the European Union as a whole and, if possible, also of the associated countries. The ELM is conceived as an online system for collecting data and providing detailed up-to-date information on the linguistic situation and its development in the various member states of the European Union and possibly, also, other European countries.[3] The project group consists of collaborators from various EFNIL member institutions. Tamás has been a great help in the software engineering and as an advisor on adapting standard software to the needs of the project. Following several surveys since 2009, the results of the latest round are now available online.[4]

4. Tamás as General Secretary of EFNIL from the inside

When Tamás was elected General Secretary in 2010, he chose Gabriella Kovács as his assistant. Gabriella was always very positive about working together with Tamás. When she changed to another institution five years later, Ivett Benyeda became the assistant and took care of the daily tasks at EFNIL for a year. At the time, Réka Dodé had been working with Tamás as project assistant for two years. He then asked her to take over the Secretariat of EFNIL in July 2016.

As Réka Dodé has known from the outset, working with Tamás was very dynamic and had room for improvement. He does not like people coming to him with every tiny problem, yet, at the same time, likes to know every little detail. Striking the balance between being counterproductive and cooperative is sometimes difficult when working with him.

The first EFNIL conference Tamás and Réka got jointly involved was in Warsaw in 2016. After the conference, their work really started. Post-conference tasks included preparing the minutes, sending out the invoices and, of course, all the other extra work that comes with that until the next conference.

In addition to Tamás many other positive attributes, empathy, attenti-

[3] See for details http://efnil.org/projects/elm
[4] See https://juniper.nytud.hu/elm4/index

veness and communication prevail in his work. Tamás pays great attention to detail and chooses his words very carefully, adapting them to the person he is talking to, both in his native Hungarian as well as in fluent English. Along with strong communication skills, his extraordinary multifunctional skills are noteworthy. He has an oustanding ability to attend to his various important projects simultaneously. He effectively outsources the odd task, but also takes his fair share of the work to be done.

All these qualities have made Tamás an ideal General Secretary of EFNIL as an international organisation. Beside the president, the Secretariat is the heart of EFNIL's life that keeps the system running. The Secretariat maintains the network between EFNIL's members. The members can turn to the Secretariat with even the slightest problem. In addition to the official email exchanges between the Executive Committee, President and the members, the Secretariat promotes the relationship and exchange of information between the various members in a friendly way. This friendly, trust-based relationship is nurtured within the EFNIL by the Secretariat and within it by Tamás. The members see Tamás as a "friend in need" who can solve any of their problems effectively. This includes a repository of relevant information that he and his collaborators offer to all members in need.

The various practical duties and tasks of the Secretariat that Tamás is responsible for and carefully performs have already been mentioned.

5. Conclusion

We are convinced that Tamás Váradi, because of his rich competence as a linguist, IT-specialist and organiser, will remain a pillar of EFNIL and its activities for the years to come. We therefore conclude our report with the traditional greeting in the old common European language of science and learning: *Ad multos annos, collega!*

References

Stickel, G., Váradi, T. (eds.) Language, Languages and New Technologies. ICT in the Service of Languages. Peter Lang, Frankfurt a.M./Berlin/Bern etc. (2011)

Stickel, G., Váradi, T. (eds.) Lexical Challenges in a Multilingual Europe. Peter Lang, Frankfurt a.M./Berlin/Bern etc. (2013)

Nyelvművelés és nyelvi tanácsadás. Vázlatos áttekintés a hetvenes évek közepétől máig

Heltainé Nagy Erzsébet[1,2]

[1] Nyelvtudományi Intézet
[2] Károli Gáspár Református Egyetem
heltaine.nagy.erzsebet@nytud.hu

1. Bevezetés

Jelen írásomban a nyelvművelés értelmezéséből kiindulva az intézeti nyelvi tanácsadó szolgálatot tekintem át, mindenekelőtt az ebben töltött évtizedek közvetlen tapasztalata alapján. Bár alapvetően kronologikus rendet követek, nem tudománytörténeti fejezet megírása a célom, hanem az, hogy a nyelvművelés pozitív hagyományát és a nyelvi tanácsadást mint a nyelvművelés jó gyakorlatát középpontba állítsam, és ezzel aktualizáljam értékeit, eredményeit, jelentőségét.

A nyelvi tanácsadásnak – mint látni fogjuk – több sikeres időszaka volt. A mostanit is annak tartom, hiszen a nyelvről való tudásanyag és a korszerű nyelvtechnológia összekapcsolásával, új eszközök alkalmazásával az utóbbi években lényegében egyenletesen nő a tanácsadás forgalma és hatékonysága. Mindez Váradi Tamás osztályvezetőnk támogatásával következhetett be, akinek ezúton, 70. születésnapján köszöntve mondok köszönetet a Nyelvművelő és nyelvi tanácsadó kutatócsoport ügyeinek elvi és gyakorlati támogatásáért.

1.1. A nyelvművelés és a nyelvi tanácsadás értelmezése

1.1.1. A nyelvművelést a nyelvhelyességi és helyesírási normák érvényesítésénél, a szorosan értelmezett nyelvtudományi megközelítésnél jóval szélesebb körű tevékenységnek tartom, beleértem a mindennapi, az amatőr, a tanári, az írói, a publicisztikai megközelítéseket, a nyelvi ismeretterjesztést, sőt az anyanyelvi nevelést és a szépirodalom szerepét is. Ha viszont a *nyelvművelés* szót szűkebb, szaktudományos értelemben használom, akkor az anyanyelvvel való tudatos törődés folyamatának nyelvtudományi megközelítését, kontrollját, a magyar nyelv iránti felelősség és elkötelezettség hagyományát, e hagyomány továbbadását értem rajta. Azt a tudatos, tipikusan értelmiségi tevékenységet, amely az anyanyelvvel, mint a kultúra hordozójával és teremtőjével, vagyis a nyelv

DOI: https://doi.org/10.18135/VT70.4

szubsztancia voltával számol – a magyar kultúrában a 16. század óta folyamatosan. Mindazzal, amit a beszélőknek, így nekem is az anyanyelv jelent: általános, mégis nagyon személyes érzéseket, identitásomat, önazonosságomat, biztonságomat, otthonomat, szűkebb és tágabb hazámat. Heidegger ismert metaforájával: *a nyelv a lét háza*, és az *ember a lét házában lakozik*. A tér és idő, a *lét háza* nekem, nekünk a magyar nyelv.

A *nekünk* itt arra utal, hogy nemcsak az egyénre, a közösségre is igaz, hogy *a nyelv a lét háza.* A magyar nyelvközösség is, jól tudjuk, *nyelvében él,* a magyar nemzet is *a maga nyelvén lett tudós,* hasonlóan más nemzetekhez. Tudjuk, ismerjük azokat az ünnepi, magasztos, olykor szakrális metaforákat is (mint a *kincs, édesanya, erős vár, tündérvár, bástya*), amelyekkel a nyelv iránti szeretetet, hűséget és felelősséget íróink, költőink fogalmazták meg, és az irodalomban, a közösségi emlékezetben hagyományozták ránk. Jókai például így, most csak őt idézem: „Egy kincse van minden nemzetnek adva, / Míg azt megőrzi híven, addig él. / E kincs neve: az édes anyanyelv".

Tudjuk azt is, hogy a magyar nyelv ügye – létünk, megmaradásunk, magyarságunk és európaiságunk ügye. Tudjuk, hogy a nyelv számunkra különleges, kiemelt érték. Volt, hogy nyelvünk a nemzeti függetlenség egyetlen fellegvára maradt. Volt, hogy megújítása európaiságunk megtartását jelentette. Az utóbbi évszázadban meg a külhoni magyarság fennmaradásának – Sütő András metaforájával szólva – egyetlen megmaradt *uszonya*. Történelmünk évszázadaiban a nyelvművelés ennek az ügynek volt a hordozója, némileg alakítója is. Mindezeket azért említem most is, hogy érzékeltessem, a nyelv ügye nálunk a nyelvészeten és az irodalmon is túl – kulturális, társadalmi, közéleti, ha tetszik politikai kérdés. Polgárosodási utak, irányok, értékrendek összeütközése minden nyelvi vita: így volt Kazinczyék korában, így a Szarvas Gábor-i új ortológia idején, így a rendszerváltozás utáni polémiákban, melyekben maga a nyelvművelés is megkérdőjeleződött. És így van ma is, amikor megint az útkeresés, a befogadás és önfeladás régi dilemmája, a hatalom birtoklásának kérdése húzódik meg az egyre hangosabb közéleti szócsatákban, vitákban.

1.1.2 A szaktudományos nyelvművelés, noha véleményem szerint mindezekkel összefügg, persze más jellegű, sokkal inkább gyakorlati tevékenység, amely elsődlegesen a mindenkori normatív nyelvhasználatra irányul. A nyelvművelés ilyen értelemben „az alkalmazott nyelvtudomány azon ága, amely a nyelvhelyesség elvei alapján, a nyelvi műveltség terjesztésével igyekszik segíteni a nyelv egészséges fejlődését"

(Grétsy és Kovalovszky szerk., 1985: 349). Ez a gyakorlat maga is nagyon összetett, és ennek magának is jelentős hagyománya van: beletartozik a nyelvi kodifikáció, a nyelvtervezés, a szókincsbővítés, a spontán vagy a tudatos nyelvújítás, a szorosan vett nyelvi és stilisztikai helyesség, a helyesírási norma. És szerves része az a nyelvi tanácsadó tevékenység, amely nálunk a nyelvtudományon belül kapott helyet, és intézményesült formájában a nyelvművelés gyakorlati területeként működik ma is.

2. A nyelvművelés gyakorlata, közönségszolgálat, később nyelvi tanácsadás

2.1. A kezdetek

A Nyelvtudományi Intézetben a kezdetektől, 1951-től alkalmilag, 1957-től pedig rendszeresen működött olyan nyelvi tanácsadó szolgálat, amelyhez közvetlenül fordulhatnak kérdéseikkel, problémáikkal a nyelvhasználók. Ahol jelezhetik a nyelvhasználatban bekövetkező, számukra feltűnő változásokat, véleményezhetik az új szavakat, és tanácsot kérhetnek helyesírási, nyelvhasználati, stilisztikai problémáikban.

A közönségszolgálat, ahogy az első évtizedekben nevezték, szervezetileg a Lőrincze Lajos vezette nyelvművelő osztály keretében indult, és a nyelvművelés részterületeként működött, Ferenczy Géza, Grétsy László, Ruzsiczky Éva, R. Lovas Gizella, Szűts László, Tompa József, T. Urbán Ilona és mások, pl. H. Molnár Ilona és Szépe György részvételével. Lőrincze Lajos, az alapító, így írt erről a kezdeti időszakról: „Az Akadémia Nyelvtudományi Intézetében állandóan cseng a telefon. Érdeklődik a középiskolás diák, aki az Akadémia tekintélyére támaszkodva szeretné megnyerni fogadását; érdeklődnek a szerkesztőségek, hogy minél jobb magyarsággal kerüljön az olvasó kezébe a magyar sajtó” (Lőrincze, 1953: 9). A közönségszolgálat ezen időszakáról l. még Ruzsiczky, 1961; Ferenczy és Ruzsiczky szerk., 1963; Lőrincze, 1968. Itt utalok rá, hogy az intézeti nyelvi tanácsadás történetéről, helyzetéről és lehetőségeiről alapos, dokumentumokra támaszkodó áttekintés található Ludányi Zsófia frissen megjelent tanulmányában (Ludányi, 2020a, különösen 330–345).

2.2. A megerősödés. Közönségszolgálat és nyelvi ismeretterjesztés

A hetvenes évek közepén, 1976-ban lettem az osztály tagja, már Grétsy László volt az osztály vezetője, az Intézet a Várban, a Szentháromság

utcai épületében volt. A közönségszolgálat ekkor és a következő évtizedben is napi rendszeres, fél 9-től fél 2-ig tartó telefonügyeletet, postai levelezést és a földszinti sarokszobákban személyes „ügyfélfogadást", megbeszéléseket jelentett. A közönségszolgálatban Grétsy Lászlóval együtt az osztály állandó és időszakos (akkori meg későbbi) munkatársai közvetve vagy közvetlenül szinte valamennyien részt vettek, emlékezetem szerint: Bíró Ágnes, Balogh Judit, Éder Zoltán, Egedy Mária, Előry Vilma, Felde Györgyi, Heltainé Nagy Erzsébet, Huszár Ágnes, Kardos Tamás, Kemény Gábor, Kovalovszky Miklós, Ladó János, Lőrincze Lajos, R. Lovas Gizella, Seregy Lajos, Tolcsvai Nagy Gábor.

Az a gazdag nyelvi anyag, amely ebben az időszakban gyűlt össze, a folyamatosan vezetett közönségszolgálati naplókban őrződött meg, és mára az elektronikus archívum történeti részét képezi. Az új szóalakok egy része közben „szótáréretté" vált, beépült az értelmező és helyesírási szótárakba, a nyelvművelő kötetekbe; folyóiratokba, főként a *Magyar Nyelvőr* és az 1979-ben induló *Édes Anyanyelvünk* ismeretterjesztő rovataiba, cikkeibe.

A nyelvművelő – akkor Mai magyar nyelvi osztály – munkatársának lenni persze nemcsak a közönségszolgálatot jelentette, hanem folyamatos kutatást, ismeretterjesztést is. A *Nyelvművelő kézikönyv* munkálatai ez idő tájt fejeződtek be (Grétsy és Kovalovszky szerk., 1980, 1985), és olyan közös munkák készültek, mint pl. *Mai magyar nyelvünk* (Grétsy szerk., 1976), *Hivatalos nyelvünk kézikönyve* (Grétsy szerk., 1978), majd *Nyelvészet és tömegkommunikáció* (Grétsy szerk., 1985), *Nyelvi divatok* (Bíró és Tolcsvai Nagy szerk., 1985), *Iratszerkesztési és fogalmazási tanácsadó* (Deme és Grétsy szerk., 1987); később a *Képes diákszótár* (Grétsy és Kemény szerk., 1992), a *Nyelvművelő kéziszótár* (Grétsy és Kemény szerk., 1996 és Grétsy és Kemény szerk., 2005).

2.3. Normativitás és nyelvművelés

Ezt az időszakot véleményem szerint a normatív nyelvhasználatra irányuló nyelvművelés (egyik) sikeres korszakának lehet nevezni. Adott volt a szilárd akadémiai intézményi háttér, a viszonylag nagy létszámban működő nyelvművelő osztály, Grétsy László vezetésével, voltak külső kommunikációs csatornák (sajtó, rádió, televízió), új nyelvművelő folyóirat *(Édes Anyanyelvünk)* indult, országos rendezvények irányították újra és újra a nyelvre, a nyelvhasználatra a figyelmet (magyar nyelv hete). Mindezek jelentős mértékben hozzájárultak a nyelvi ismeretterjesztés presztízséhez, a nyelv iránti érdeklődés fenntartásához.

Céljából és jellegéből adódóan ez a nyelvművelő tevékenység a normativitás, a nyelvi egységesülés folyamatát hangsúlyozta, de tágabb értelemben azt a kulturális hagyományt vitte tovább, amely a nyelvet kiemelt értéknek tekintette a közösségi, a nemzeti művelődés fenntartása szempontjából, a nyelvvel való törődést pedig tudatos, értelmiségi magatartásnak, kulturális hagyománynak tekintette.

Kitüntetett normája a köznyelv, tegyük hozzá, az igényes, akkori szóhasználattal a művelt köznyelv volt, amelynek írott és hangzó rétegei a nyilvánosság előtti megszólalás fontos területeinek számítottak, mint a sajtó, rádió, televízió, színház, vagy mint az iskola, az anyanyelvi nevelés fontos színtere. Ez a nyelvművelés jellegéből adódóan nyilvánvalóan a köznyelvet beszélő rétegeket hozta előnyösebb helyzetbe, de összességében semmiképpen sem volt ellenséges a nyelvjárási vagy a regionális nyelvi rétegekkel, főként nem a beszélőkkel. Sőt! Az ismeretterjesztésben, rádió-előadásokban, cikkekben, tanulmányokban a nyelvjárások, a nyelvjárási szavak és fordulatok – a klasszikus szépirodalom mellett – mint fontos normaképző összetevők jelentek meg. Gondoljunk csak Lőrincze Lajos, később Szathmári István ízes beszédű rádiós ötperceire, vagy Grétsy László színes, élő nyelvi ismeretterjesztő műsoraira. (Azt most csak zárójelben jegyzem meg, hogy „a nyelvművelők" maguk is többen részt vettek a nyelvjárások, a nyelvjárási szókincs felgyűjtésében, dokumentálásában, és gyakran maguk is egy-egy nyelvjárási régióban voltak otthon.)

3. A nyelvművelés gyakorlata, nyelvi tanácsadás

3.1. Útkeresések, változások

A fenti stabil évtized után a megváltozó külső és belső körülmények hatására változások következtek be a nyelvművelő osztály életében. Az osztály megnevezése, struktúrája többször változott, volt Nyelvművelő osztály, Mai magyar nyelvi osztály, később Normatív nyelvi osztály, majd újra Nyelvművelő osztály. A Petőfi-szótár elkészülte után J. Soltész Katalin és Wacha Imre átmenetileg, Előry Vilma tartósan az osztály munkatársa lett. Grétsy László 1987-ben távozott az Intézetből. 1988 és 1993 között Kemény Gábor volt az osztály vezetője, a kutatások ekkoriban újdonságként a nyelvi norma területeire irányultak (l. pl. *Normatudat – nyelvi norma*, Kemény szerk., 1992). Később Szűts László lett az osztályvezető, 2003-tól pedig Posgay Ildikó. A létszám a kilencvenes években stagnált, utána csökkent, átmenetileg fiatal kutatóval, Domon-

kosi Ágnessel bővült. Az Intézet közben kétszer is költözött: a Szentháromság utcából a Színház utcába, a Várszínház épületébe, majd mai helyére, a Benczúr utcába. A közönségszolgálat ebben az időszakban is folyamatosan működött.

Változtak a külső körülmények is. A 90-es években kibontakozó szakmai, még inkább közéleti vitákban az a konszenzus, amely a nyelvművelés pozitív kulturális és közéleti tartalmát és hagyományát jelentette, mint a bevezetőben utaltam rá, megkérdőjeleződött, eredményei, értékei viták kereszttüzébe kerültek. Bár kétségtelen, hogy a viták hozzájárultak az alapkérdések újragondolásához, melyre különösen a határokon túli, a külhoni magyar nyelvközösségek szempontjából volt égető szükség, a nyelvművelés, a nyelvvel való tudatos törődés, a nyelvi folyamatok értékelésének tekintélye és hagyománya mélyen leértékelődött. Természetes, hogy mindez rányomta bélyegét a nyelvművelő osztály helyzetére is.

A kétezres évek elején az akadémiai hátterű határon túli nyelvi irodák hálózata (mai nevén Termini Magyar Nyelvi Kutatóhálózat)[1] és az intézeti nyelvi tanácsadó szolgálat szerves kapcsolatot alakíthatott ki. Különösen a dunaszerdahelyi Gramma Nyelvi Irodával volt hatékony az együttműködés, amelynek keretében megindultak, folytatódtak a gyakorlati nyelvművelést, a nyelvi tanácsadást, a nyelvi menedzselést középpontba állító kisebb konferenciák és máig emlékezetes műhelytalálkozók, szakmai beszélgetések formájában. Az együttműködés eredménye a *Műhelytanulmányok a nyelvművelésről* (Domonkosi és mtsai. szerk., 2007), amelyben a szerzők a nyelvművelés, a nyelvtervezés elméleti kérdésköreit és a nyelvi szolgáltatások gyakorlatát tekintik át egy- és kétnyelvű környezetben, foglalkoznak a stílusminősítésekkel, helyesírási kodifikációs problémákkal, a nyelvi babonák, mítoszok jellegével, szerepével, a purizmus és az anyanyelvi nevelés kapcsolatával.

3.2. Nyelvi tanácsadás – új nyelvtechnológiai környezetben

A kétezres évekre jelentősen megváltoztak a külső technikai körülmények is, a közönségszolgálatból a megújulást nevében is jelző *nyelvi tanácsadó szolgálat* lett. A munkatársi létszám tovább csökkent, a telefonszolgálatot, a növekvő, és már egyre inkább elektronikus levelezést, valamint az egyéb teendőket a munkatársak már csak csökkentett mértékben és időtartamban tudták ellátni.

[1] http://termini.nytud.hu

Ez idő tájt az Intézetben újabb, szerkezeti, szervezeti változások történtek. 2010-ben az Élőnyelvi osztály, a Nyelvművelő osztály és a Nyelvtechnológiai osztály egyesült, amelyből azután létrejött a Váradi Tamás vezette mai Nyelvtechnológiai és Alkalmazott Nyelvészeti Osztály. Ebben a helyzetben a tanácsadás működése, akkori jelene és jövője szempontjából szerencsés volt, hogy a Nyelvtechnológiai és Alkalmazott Nyelvészeti Osztály része lett. A tanácsadó szolgálatnak ez a változás új lehetőséget teremtett a technikai és szakmai fejlesztésre, a korszerűsítésre. A *Nyelvművelő és nyelvi tanácsadó kutatócsoport* mint az osztály egyik autonóm kutatócsoportja, kibővített lehetőségekkel, a korszerű nyelvtechnológiai háttér támogatásával meg tudott maradni, és el tudta látni feladatait – a folyamatosan növekvő érdeklődés mellett is – eredményesen, sikeresen. Itt mondok köszönetet az osztály tagjainak a helyesírási weboldal készítésekor a közönségszolgálati adatbázis archívummá alakításában vagy az Utónévkereső létrejöttében nyújtott nyelvtechnológiai támogatásukért. Külön örömömre szolgál, hogy Ludányi Zsófia közvetlen munkatársunk, csoporttagunk lett, és folyamatosan részt vesz a nyelvi tanácsadásban és a nyelvi ismeretterjesztésben. Nyelvtechnológiai ismeretei pedig anyagaink rendezésében, archiválásában, a technikai háttér működtetésében nélkülözhetetlenek.

A Nyelvművelő és nyelvi tanácsadó kutatócsoport az utóbbi időben (jelenleg is) négy munkatársat jelent: Heltainé Nagy Erzsébet, Ludányi Zsófia, Kardos Tamás a nyelvi tanácsadó szolgálat feladatait látják el, Raátz Judit (félállásban) az utónévügyekkel kapcsolatos teendőket. A nyelvi és a névügyi tanácsadás, szakvéleményezés, szövegellenőrzés, a telefonszolgálat, az egyre növekvő levelezés teljes mértékben lekötik a csoport kapacitását, ezért, már a közeljövőre gondolván is, csoportvezetőként feltétlenül szükségesnek tartom a munkatársi létszám növelését, fiatal kutató(k) bevonását, természetesen a technikai, nyelvtechnológiai fejlesztésekkel párhuzamosan.

3.3. A megújult nyelvi tanácsadás jellege és funkciói

A nyelvi változások, a megjelenő új szavak, változatok helyes, a szabályoknak megfelelő írásmódja, a nyelvhelyesség kérdései, a sztenderd normának való megfelelés a nyilvános színtereken – ezekben a címszavakban foglalhatók össze ma is a nyelvhasználókat leginkább foglalkoztató nyelvi kérdések. A hozzánk fordulók továbbra is kifejezhetik véleményüket, tudathatják az általuk észlelt nyelvi változásokat, tanácsot,

megerősítést kérhetnek az egyes nyelvi formák használatával kapcsolatban. Ez a lehetőség a tanácsadás legfőbb funkciója ma is.

A kérdezők motivációja, a nyelvi kérdések iránti érdeklődés ma is szilárd, az intézményesült nyelvi tanácsadás a tudatos nyelvhasználók igényéből adódóan működik napjainkban is. Azok fordulnak hozzánk, akik a kifejezőkészséget, a nyelvtani helyességet és a választékosságot, a helyesírási készséget és tudást fontosnak tartják, akiknek a munkájukhoz szükségük van ilyen tudásra. Cégek, vállalatok marketingfelelősei, reklámszakemberei, állami és közintézmények, kormányhivatalok dolgozói, a sajtó munkatársai, szerkesztők, lektorok, tanárok, diákok, magánemberek – állandó és újabb partnereink a tanácsadó szolgálatban.

A csoport munkatársai a tanacs@nytud.hu címen fogadják a nyelvhasználattal kapcsolatos kérdéseket, de van lehetőség közvetlen, telefonos érdeklődésre is. A kérdések szakszerű megválaszolása alapvetően tanácsadás: a lehetőségek, a változatok, a nyelvi variabilitás bemutatása, a normatív nyelvhasználatra érvényes következtetések feltárása, javaslatok, ajánlások megfogalmazása. A nyelvi tanácsadás mint alkalmazott nyelvészeti tevékenység, a nyelvtudomány számos területére közvetlenül épít, ilyen a nyelvleírás, a történeti nyelvészet, a lexikográfia, a stilisztika, a retorika, a névtan, a korpusznyelvészet.

További fontos funkciója a tanácsadásnak ma is a nyelvi ismeretterjesztés, a nyelvtudomány újabb eredményeinek és eszközeinek megismertetése. A nyelvről szerzett ismeretek hozzájárulnak a nyelvi attitűdjelenségek, a pozitív és negatív vélekedések differenciálódásához, a nyelvhasználattal kapcsolatos szemléletváltozáshoz.

A Nyelvtudományi Intézetben működő nyelvi tanácsadás jelentős szerepet tölt be a tudomány és a beszélőközösségek kapcsolatának fenntartásában. Továbbviszi azt a hagyományt, melyet az akadémiai, az intézményesült nyelvművelés, a tudatos nyelvgondozás az elmúlt évszázadokban betöltött. Különösen igaz ez az olyan területen, mint a helyesírási kodifikáció, az írásgyakorlat gondozása és egységének fenntartása.

Ezt a területet azért is említem, mert a kérdések jelentős része ma is helyesírási, köztük is vezetnek az új átvételek, új nyelvi formák (új dolgok, jelenségek, tárgyak) írásmódjával kapcsolatos kérdések. A legutóbbi időben természetesen a világjárvánnyal, a vírushelyzettel összefüggésben megjelent új átvételek, köznyelvivé vált szakszavak, régi-új kifejezések, jelzős szerkezetek, összetételek írásmódja kérdéses, például: *Covid19, digitális tanítás, hibrid oktatás, home office, kijárási korláto-*

zás, *koronavírusteszt, nyájimmunitás, oltakozás, online színház, pandémia, vírustagadás* stb. (Bővebben a témáról l. még: Ludányi, 2020b.) A nyelvi tanácsadás területeinek, a tanácsadás gyakorlatának részletesebb bemutatására, értékelési szempontjaira pedig Heltainé Nagy, 2012, 2013.

3.4. A nyelvtechnológiai osztály és a tanácsadó szolgálat

Az együttműködésnek a nyelvi tanácsadás szempontjából is fontos eredménye a már említett, 2013 óta működő, a felhasználók körében közkedvelt *helyesírás.mta.hu* weboldal. A Váradi Tamás vezette nyelvtechnológiai kutatócsoport által kidolgozott, teljesen automatizált rendszer (bemutatására l. pl. Váradi, 2013, Váradi és mtsai., 2014) szoros kapcsolatban van a nyelvi tanácsadó szolgálattal is, hiszen ahol erre szükség mutatkozik közvetlenül „visszavezeti" a felhasználót a személyes konzultációhoz, a tanácsadóhoz vagy a tanacs@nytud.hu címhez.

A nyelvi tanácsadás gyakorlata felől is jól érzékelhető, hogy az eszköz látványosan megnövelte a nyelvi tanácsadás ismertségét, hatékonyságát. Folyamatosan nőtt a levelezés mértéke, bár nem kiugró mértékben, de összességében növekszik a kérdések és válaszok száma: 2013-ban 814, 2014-ben 906, 2015-ben 858, 2016-ban 917, 2017-ben 1030, 2018-ban 1128, 2019-ben 999 és 2020-ban 1132.[2] Tehát a weboldal nem hogy „nem vette el" a tanácsadók munkáját, hanem támogatta, segítette. A kérdezők és a kérdések számát, a nyelvi ismeretek iránti érdeklődést – pedig megnövelte.

3.5 A kutatócsoport névtani feladatai és az Utónévkereső

Mivel a kutatócsoporthoz tartozik az anyakönyvezhető keresztnevek névtani szempontú véleményezése és a névanyag gondozása, fontosnak tartom megemlíteni a nyelvtechnológiai és a nyelvművelő csoport másik közös fejlesztését, az *utónévkereső portált*,[3] amely a Magyarországon anyakönyvezhető keresztnevek anyagában jelent sokféle tájékozódási lehetőséget. Az anyakönyvezhető férfi és női nevek aktualizált, folyamatosan bővülő listáján túl az Utónévkereső névválasztási szempontokkal is foglalkozik. A funkcióiban folyamatosan bővülő portálon figyelemmel kísérhetik a látogatók a nevek számát, a pontos névalakokat, a legfris-

[2] A statisztikát Ludányi Zsófiának köszönöm.
[3] http://corpus.nytud.hu/utonevportal

sebb névadási trendeket, névválasztáshoz kaphatnak keresési és felhasználási lehetőséget. A személyes érdeklődés és tanácsadás lehetősége ezen a területen is megmaradt (a nevtanacs@nytud.hu címen).

4. Összegzés, kitekintés

A nyelvi tanácsadásnak több sikeres időszaka volt. A mostani is az, hiszen a nyelvről való tudásanyag és a korszerű nyelvtechnológia összekapcsolásával, új eszközök alkalmazásával megnőtt a tanácsadás hatékonysága. A személyes tanácsadás lehetősége is minden internetes technikai eszközön elérhető: számítógépen, laptopon, tableten, okostelefonon; levelezésben és közvetlen telefonhívásokon keresztül is. A letöltések és a látogatók száma pedig mérhető, és mindkettő folyamatos növekedést mutat.

Hogy mindezzel arányosan nőtt-e az ismeretek megszerzése, felhasználása, alkalmazása? A visszajelzések alapján kérdezőink esetében – igen. Változnak-e a nyelvhasználattal kapcsolatos vélekedések? Erre még nem látjuk a választ. Remélhetőleg: igen. De az biztos, hogy a technológia felhasználása, a nyelvi tanácsadás korszerűsítése lehetséges és szükséges háttere a nyelvi tájékozódásnak is. Nyilvánvalóan nem helyettesíti az iskolai és az otthoni anyanyelvi nevelés szerepét, de jelentheti a további anyanyelvi művelődés lehetőségét, akár iskolai eszközként való felhasználását, ezáltal is ébren tartva a nyelv és a nyelvtudomány iránti érdeklődést.

A nyelv ügye – visszautalva bevezető gondolataimra – mindezeken túl ma is kulturális, társadalmi, közéleti kérdés. Anyanyelvünk ügye létünk, magyarságunk és európaiságunk ügye ma is. A nyelvművelés, melynek része a nyelvi tanácsadás, ennek az ügynek a hordozója, és némileg talán alakítója lehet ma is. Akár olyan „apróságokon” át, mint egy-egy helyesírási, nyelvhelyességi, stiláris kérdés és válasz. Mert az is, ez is felelősség. „Az íróság ott kezdődik, mikor az ember felelősséget érez egy alany és egy állítmány összefűzésekor is; mert az is becsületvizsga: állítás, amiért helyt kell állni. Egy életen át! Egy nemzet életén át!” – fogalmazta meg így Illyés Gyula a maga írói „ráérzésében” (1975: 16).

Bibliográfia

Bíró Á., Tolcsvai Nagy G. (szerk.) Nyelvi divatok. Gondolat Könyvkiadó, Budapest (1985)

Deme L., Grétsy L. (szerk.) Iratszerkesztési és -fogalmazási tanácsadó. Közgazdasági és Jogi Könyvkiadó, Budapest (1987)

Domonkosi Á., Lanstyák I., Posgay I. (szerk.) Műhelytanulmányok a nyelvművelésről. Tinta Könyvkiadó, Fórum Kisebbségkutató Intézet, Gramma Nyelvi Iroda, Budapest, Dunaszerdahely (2007)

Ferenczy G., Ruzsiczky É. (szerk.) Nyelvművelő levelek. Az Akadémia Nyelvtudományi Intézetének levelesládájából. Gondolat Könyvkiadó, Budapest (1964)

Grétsy L. (szerk.) Mai magyar nyelvünk. Akadémiai Kiadó, Budapest (1976)

Grétsy L. (szerk.) Hivatalos nyelvünk kézikönyve. Pénzügyminisztérium Államigazgatási Szervezési Intézet, Budapest (1978)

Grétsy L. (szerk.) Nyelvészet és tömegkommunikáció I–II. Membrán könyvek. Tömegkommunikációs Kutatóközpont, Budapest (1985)

Grétsy L., Kemény G. (szerk.) Képes diákszótár. Akadémiai Kiadó, Budapest (1992)

Grétsy L., Kemény G. (szerk.) Nyelvművelő kéziszótár. Auktor Kiadó, Budapest (1996)

Grétsy L., Kemény G. (szerk.) Nyelvművelő kéziszótár. Tinta Könyvkiadó, Budapest (2005)

Grétsy L., Kovalovszky M. (szerk.) Nyelvművelő kézikönyv I-II. Akadémiai Kiadó, Budapest (1980, 1985)

Heltainé Nagy E.: Nyelvhasználati minősítések, a helyes-helytelen a tanácsadói gyakorlatban. Magyar Nyelvőr 136/4, 394–406 (2012)

Heltainé Nagy E.: A nyelvi tanácsadás területei és újabb eszközei az MTA Nyelvtudományi Intézetében. Anyanyelv-pedagógia 7/1. (2013) http://www.anyanyelv-pedagogia.hu/cikkek.php?id=505

Illyés Gy.: Anyanyelvünk. Magvető Könyvkiadó, Budapest (1975)

Lőrincze L.: Nyelv és élet. Művelt Nép Könyvkiadó, Budapest (1953)

Lőrincze L.: Nyelvőrségen. Akadémiai Kiadó, Budapest (1968).

Ludányi Zs.: Nyelvi menedzselés és nyelvi tanácsadás. Helyzetkép, lehetőségek, feladatok. Magyar Nyelvőr 144/3, 318–345 (2020a)

Ludányi Zs.: Helyesírási kérdések a pandémia idején. Magyar Szó Online (Novi Sad). https://www.magyarszo.rs/hu/4391/mellekletek_uveggolyo/225543/Helyes%C3%ADr%C3%A1si-k%C3%A9rd%C3%A9sek-a-pand%C3%A9mia-idej%C3%A9n.htm (2020b)

Ruzsiczky É.: Egy év a telefon mellett. Magyar Nyelvőr 85/2, 170–184 (1961)

Váradi T.: helyesírás.mta.hu – Helyesírási tanácsadó portál. Édes Anyanyelvünk 35/4, 17 (2013)

Váradi T., Ludányi Zs., Kovács R.: Géppel segített helyesírás. A helyesírás.mta.hu portál készítéséről. Modern Nyelvoktatás 20/1–2, 43–58 (2014)

A Magyar nemzeti szövegtár

Oravecz Csaba[1]

[1]Westpole Luxembourg
oravecz.csaba@gmail.com

1. A kezdetek[1]

A számítógépek megjelenésével szinte egy időben felmerült azok alkalmazása a nyelv elemzésére. Már a 1960-as években, de különösen a 80-as évektől a számítógépek tömeges elterjedésével új távlatok nyíltak a nyelv empirikus vizsgálata terén. Egy új ága született a nyelvészetnek, a korpusznyelvészet, amely a nyelvhasználat számítógépes modellezését tűzte ki célul. Ehhez nagy méretű szöveges adatbázisokat (korpuszokat) építettek, amelyek egy időben és térben meghatározott közösség nyelvhasználatának nyelvileg elemzett reprezentatív mintáját jelentik. Alapvetően a Nyelvtudományi Intézetben Váradi Tamás vezetésével 1997-ben létrejött, akkor még Korpusznyelvészetinek nevezett osztály is egy ilyen adatbázis, a Magyar nemzeti szövegtár (MNSz.) megépítését tűzte ki célul. Ez a kezdeményezés akkoriban nálunk mind módszertanában, mind volumenében újdonságnak számított (ez sokáig az osztályon dolgozó kevesek létszámában is megmutatkozott), de talán nem túlzás azt állítani, hogy a magyar nyelvre irányuló adatközpontú számítógépes nyelvészeti kutatásoknak és a nyelvfeldolgozó alkalmazások fejlesztésének egyik megkerülhetetlen kiindulópontjává vált. Ma már persze természetes a nagy mennyiségű nyelvi adat nélkülözhetetlensége, és a felhasználására vonatkozó egyre növekvő igény, de mintegy negyedszázaddal ezelőtt egy ilyen vállalkozás sok mindenben úttörőnek számított magyar nyelvterületen.

Az adatközpontú módszerek és alkalmazások fejlődése az utóbbi időben még inkább szembetűnő. A számítógépek teljesítményének növekedése, a szövegadatbázisokból, korpuszokból automatikusan tanulni képes, gépi tanuló eljárások kifejlesztése, a neurális hálózatok térhódítása

[1] Ez a dolgozat a Magyar Tudományban 2014-ben megjelent Váradi Tamás – Oravecz Csaba „A Magyar Nemzeti Szövegtár egymilliárd szavas új változata” című tanulmány alapján készült, abból kisebb változtatásokkal vagy anélkül hosszabb szövegrészeket is tartalmaz.

DOI: https://doi.org/10.18135/VT70.5

a számítógépes nyelvészet, a nyelvtechnológia intenzív fejlődéséhez vezetett, és tovább szélesíti, alakítja át a(z elméleti) nyelvészeti kutatások spektrumát és módszertanát is.

A számítógéppel végzett, illetve segített nyelvészeti kutatások, a nyelvtechnológiai alkalmazások alapvető feltétele a naprakész, a nyelvhasználatot reprezentatív módon tükröző, géppel olvasható és feldolgozható nyelvi adat, mely minden elméleti és alkalmazott kutatás kiindulópontja, a nyelvtechnológiai alkalmazások fejlesztésének nélkülözhetetlen nyersanyaga. Az ezeket az adatokat tartalmazó nyelvi adatbázisok pontos, számszerűsíthető képet adnak a nyelvhasználatról, egyben megkerülhetetlen forrásai és bemeneti adatai nyelvfeldolgozó algoritmusoknak, valamint értékes információt hordoznak az adott nyelvhez kötődő kultúra kutatóinak, társadalomtudósainak számára is.

A Szövegtár első változata 1998 és 2001 között készült, és a 90-es évek második felének nyelvhasználatából merített reprezentatív mintával a magyar nyelv első, az akkori gyakorlatban is jelentős méretűnek számító, nyelvileg elemzett korpusza volt, amely hálózati lekérdező felületen bárki számára szabadon hozzáférhető volt (Váradi, 2002). A munkálatok kezdetétől számított, lassan 15 év múltával vált nyilvánvalóvá, hogy általában a számítógépes korpuszokkal, így az MNSz.-szel szemben támasztott igények jelentős mértékben változtak és több szempontból megnövekedtek, különösen az alábbi 3 területen:

- **Minőség.** A számítógépes nyelvészeti technológia gyors fejlődése miatt az MNSz.-ben alkalmazott számítógépes nyelvi elemzés technológiája, pontossága és a nyelvi információ (re)prezentációjának módszere elmaradt a nemzetközi sztenderdnek tekinthető szinttől.
- **Terjedelem.** Az első változatban előirányzott 100 millió szavas terjedelem már nem volt jelentősnek tekinthető. Az adatközpontú módszerek/alkalmazások elterjedése és sikeressége a számítógépes nyelvfeldolgozás területén a nemzetközi gyakorlatban kívánatossá tették a milliárd szavas nagyságrendű korpuszok kifejlesztését (Parker et al., 2011), mivel az adatok ugrásszerű növekedése a rajtuk alapuló alkalmazások minőségének javulását vonja (vonta) maga után.
- **Reprezentativitás, lefedettség.** A nyelvhasználat pontos, akár a nyelvtörténeti kutatások igényeit is kielégítő dokumentálása egy-

részt újabb és újabb állapotfelvételt (adatgyűjtést) igényel, másrészt a nyelvi változatok széles skáláját kell, hogy képviselje. Ebből a szempontból például az MNSz. kritikus hiányossága volt a beszélt nyelvi adatok teljes hiánya.

2. Az MNSz. előzményei és első változata

Külföldi kutatások eredményeként már a 60-as évektől rendelkezésre állnak mai mértékkel természetesen kis méretűnek számító, de gondosan összeállított korpuszok (lásd Brown-korpusz [Kučera és Francis, 1967]). A 90-es évek jelentős produktuma az MNSz.-nek is néhány szempontból mintául szolgáló British National Corpus, és ettől az időszaktól folyamatosan készültek további nemzeti korpuszok. A nagy méretű korpuszokban reprezentált nyelvi információ gépi előállítására irányuló kutatás gyakorlatilag egy külön számítógépes nyelvészeti „iparág", a különféle annotáló és egyértelműsítő rendszerek fejlesztésének kialakulásához vezetett. Ebbe a sorba illeszkedett az MNSz. első változata is, ami az ilyen nagyságrendű korpuszokhoz hasonlóan automatikus morfoszintaktikai annotációt kapott a Nyelvtudományi Intézetben kifejlesztett nemzetközi szinten is élvonalbeli pontossággal működő eljárás segítségével (Oravecz és Dienes, 2002).

Az MNSz. első változatának elkészülte óta mind a korpuszok mérete, mind az alkalmazott gépi feldolgozás minősége és részletessége megváltozott. Ma már az elvárt kategóriát jelentik a több száz millió vagy több milliárd szavas adatbázisok, és ebben a nagyságrendben az MNsz.-szel párhuzamosan elkészült a magyar Webkorpusz is (Halácsy et al., 2003). A feldolgozás tekintetében hatékonyabb, pontosabb és részletesebb nyelvi elemzést adó eljárások, alkalmazások kifejlesztését célzó kutatások szintén a 2000-es évek elején-közepén kezdődtek meg magyar nyelvre is (Halácsy et al., 2006, 2007; Trón et al., 2005).

A jelentős méretű korpuszokban tárolt nyelvi elemzés részletessége automatikus annotáció esetén általában a morfológia szintjén maradt, szintaktikailag elemzett, magyar nyelven is létező adatbázisok az elfogadható elemzési pontosság érdekében (géppel segített) kézi annotációval készültek (Csendes et al., 2004).

Napjaink milliárd szavas szövegkorpuszai ún. opportunista összeállítással, általában a weben elérhető szövegek teljes letöltésével készülnek, azaz összetételükben kevésbé törekednek a nyelvhasználat különféle változatainak kiegyensúlyozott reprezentálására.

3. A továbblépés

Az MNSz. első változata igen sikeres nyelvi erőforrásnak volt tekinthető. A Kárpát-medencei Magyar Nyelvi Korpusz projekt keretében 2005 novemberére a határon túli nyelvváltozatokkal 187 millió szóra kibővült korpusznak több ezer regisztrált felhasználója volt, az MNSz.-ben található nyelvi adatok alapján több tucat tanulmány készült. Mindezek ellenére kétségtelen, hogy a mintegy 15 év elteltével az első változat elavulttá vált.

Az új változat (MNSz.^{2}) kifejlesztésének célja az 1. fejezetben említett hiányosságok kiküszöbölésével olyan magas minőségű, megnövelt és lefedettségét illetően kibővített komplex nyelvi adatbázis létrehozása volt, amely hatékonyan képes kiszolgálni a felhasználók és kutatók megnövekedett igényeit. Ennek érdekében a fenti felosztás szerint a új változattal kapcsolatos célkitűzések az alábbiakban foglalhatók össze:

- **Minőség.** A korpusz anyagának minden feldolgozási és elemzési lépésében új, korszerű számítógépes nyelvészeti technológia felhasználása a legújabb vonatkozó fejlesztéseinek figyelembevételével és a magyar nyelvre való alkalmazásukra irányuló célzott kutatással.
- **Terjedelem.** A korpusz anyagának bővítése minimum 1000 millió szóra.
- **Reprezentativitás, lefedettség.** Újabb mintavétel a mai magyar nyelvhasználatnak a Szövegtárban addig is szereplő, valamint további változataiból. Jelentős hozzáadott értékként jelent meg a beszélt nyelvi megnyilatkozások lejegyzett formátumát tartalmazó korpuszrész kialakítása, valamint mintavétel a közösségi média szövegeiből.

4. Az új változat fejlesztése

Az MNSz.^{2} esetében az MNSz. első változatában alkalmazott technológia minden részletében felülvizsgálatra, átdolgozásra, továbbfejlesztésre került a nemzetközi eredmények és a magyar nyelvre irányuló újabb kutatások alkalmazásával. Ez a munka a korpuszépítés minden fázisában jelentkezett.

4.1. Az anyaggyűjtés

Szöveges adatok összegyűjtésére ebben a nagyságrendben a kézenfekvő módszer vagy az internet bizonyos tartományainak végigpásztázása és

az ott talált anyagok valamilyen heurisztikus szűréssel segített, de alapjában véve válogatás nélküli letöltése, vagy nagy mennyiségű sajtóanyag beszerzése. Kizárólagos alkalmazás esetén mindkét módszernek vannak egyértelmű hiányosságai, ha a cél egy kiegyensúlyozott, elegendő metaadattal ellátott korpusz összeállítása. Előbbi módszer a szűrés ellenére is gyakran nagyon zajos adatot eredményez, melyhez jellemzően az az alapvető bibliográfiai információ is hiányzik, amely nélkül alapos nyelvészeti kutatások sokszor nemigen végezhetők. Az utóbbi módszerrel előálló korpusznak a reprezentativitás hiánya a szembetűnő hátránya.

Ezért jelentős munkát kellett fordítani a korpusz anyagának kontrollált és az adott forráshoz illeszkedő begyűjtésére: a közösségi médiából származó szövegek automatikus monitorozására, számítógéppel feldolgozható és metaadatolható eredményt adó letöltésére, a különböző forrásgazdákkal történő megegyezésre az általuk birtokolt anyagok archívumához való hozzáféréshez. Azok a források, melyek már alapesetben valamilyen (félig) strukturált, jól feldolgozható formátumban álltak rendelkezésre, előnyt élveztek a vegyes formátumú esetleges összeállítású archívumokkal szemben. A gyűjtés nagyságrendje természetesen eleve kizárta a kézi beavatkozást és a nagyon zajos kimenetet adó módszereket, mint a dokumentumok szkennelése, illetve optikai karakterfelismerést igénylő dokumentumok felhasználása. Az a manuális munkaerő, amely ezeket a módszereket alkalmazhatóvá tette volna, messze nem állt rendelkezésre.

Az anyaggyűjtés során elkerülhetetlenül szembesülni kellett az utóbbi időben egyre nagyobb hangsúlyt kapó szerzői jogokkal kapcsolatos kérdésekkel. Ekkora nagyságrendben lehetetlen vállalkozás volt minden adatgazdától (ha egyáltalán beazonosítható és megtalálható) a lehető legszabadabb felhasználói jogok megszerzése. Az MNSz.^{2} így alapesetben továbbra is egy felhasználói felületen férhető hozzá.

Az az előzetes várakozás, hogy a 15 évvel ezelőtti helyzethez képest a szöveges dokumentumok kezelése és tárolása a nemzetközi szabványokhoz közelítve sokat javult, és ez majd nagyban megkönnyíti a korpusz anyagának összegyűjtését, nem igazolódott be; sok probléma adódott a forrásszövegek hozzáférhetőségével és eredeti formátumával. Ehhez adódott még egy sajnálatos további hátráltató tényező: számos olyan adatforrás, amelyeknek a szövegei az MNSz. első változatának szerves részét alkotják, nem járult hozzá az azóta keletkezett szövegeik felvételéhez az MNSz.^{2}-be. Ennek valódi okait csak találgatni lehet, szomorú következménye viszont az, hogy a nyelvhasználat bizonyos jelentős szegmentumai az MNSz.^{2} mintavételéből teljesen kimaradtak.

A korpusz végül mintegy 1,5 milliárd szóra bővült. A sajtónyelvi anyag továbbra is domináns maradt, viszont minden nyelvváltozat anyaga minimum megduplázódott a korábbi változathoz képest, valamint megjelent egy új „műfaj", a(z átírt) beszélt nyelvi anyag is.

4.2. Előfeldolgozás és szövegnormalizálás

Az előfeldolgozás és normalizálás során a cél a forrásszövegek olyan szabványos elektronikus formátumba alakítása volt, mely hatékonyan feldolgozható bemenetként szolgálhat a nyelvi elemzőlánc számára. Ebben a lépésben történik a forrásformátumokból a hasznos szöveges tartalom kinyerése és az alapvető dokumentumstruktúra azonosítása, a karakterek normalizálása. A későbbi feldolgozás szempontjából fontos lépés a nyelvazonosítás, a nem magyar nyelvű szövegrészek kiszűrése, illetve megjelölése.

A gondos forrásválogatás ellenére a szövegek között mindig megjelennek (közel) duplikátumok. Ezek detektálása az MNSz.[2] esetében annál komplexebb kérdésnek bizonyult, hogy például egy, az internetről letöltött szövegeken alapuló korpuszokra kifejlesztett sztenderd megoldást közvetlenül alkalmazni lehessen (Pomikalek, 2011). A források változatossága (a közösségi média letöltött szövegeitől a hivatalos, jogi anyagokon keresztül a sajtószövegekig és a szépirodalomig) célzott módszer alkalmazását tette szükségessé, ami egy általános eszközkészleten alapult (Kupietz, 2005), de az egyes szövegtípusokra szabott automatikus detektálást manuális ellenőrzésnek is kellett követnie, hogy megállapíthassuk, vajon valódi duplikátumokról van-e szó, vagy olyan ismétlődő szövegegységekről, melyek szerves tulajdonsága az ismétlődés, így adattorzítást éppen az eltávolításuk okozott volna (lásd például az időjárásjelentések szövegei).

4.3. Elemzés és annotáció

Az MNSz.[2] fejlesztése a nyelvi feldolgozás minden szintjén jelentős minőségi javulást eredményező új, illetve továbbfejlesztett eszközöket használt fel, többek között új automatikus egyértelműsítő architektúrát, illetve a kapott morfoszintaktikai elemzést reprezentáló új annotációs formátumot. Elsősorban a morfo(fono)lógiai és szintaktikai kutatások későbbi igényeinek figyelembevételével megvalósult a legkisebb azonosított alkotóelemek, az egyes morfémák reprezentálása, a főnévi csoportok és névelemek azonosítása; ezek az információk az MNSz.-ben még

nem voltak jelen, és ma is ritkaságnak számít ilyen méretű korpuszban a nyelvi információ ezen részletessége.

A hasznos szöveganyag nyelvi elemzésének előkészítő lépéseit (mondatokra, illetve szó jellegű elemekre bontás – szegmentálás/tokenizálás) a Huntoken eszköz továbbfejlesztett, „háziasított" változata végezte (Miháczi et al., 2003). A morfológiai elemzést, mely gazdag morfológiával rendelkező nyelvekre kritikus fontosságú a további magasabb szintű elemzéshez, a jelentősen felújított Humor morfológiai elemző (Prószéky és Tihanyi, 1996) szolgáltatta, információt adva a szótővel, egyes morfémákkal, szóösszetételekkel kapcsolatban.

A belső annotációs formátum kiindulópontja a mondatra bontás és a tokenizálás kimenete. Minden szóelem *(token)* külön sorban szerepel, üres sorok jelölik a mondathatárokat. Minden további nyelvi annotáció típusonként egy-egy újabb oszlopban jelenik meg, egy rugalmas és könnyen feldogozható formátumot eredményezve. A több szóelemen átnyúló szerkezeteket az ún. IOB formátum szerinti kódolás[2] reprezentálja. Ez a belső reprezentáció egyszerűen átalakítható szabványos XML-formátumra, amennyiben szükséges.

5. Közzététel

Az adatbázis kialakításának utolsó lépéseként a megnövelt terjedelem igényelte az adatbázist építő rendszer továbbfejlesztését is. A megnövekedett felhasználói igények kiszolgálására az MNSz.[2] teljesen új hálózati felületet kapott, a lekérdezések beépített elemzését és több szempontú rendezését segítő korszerű webes technológiát kihasználó segédeszközökkel. A felület lehetőséget ad összetett menüvezérelt keresésre a kódolt információ minden részletében. A megjelenítési beállításokban a szövegkörnyezet, a metaadatok prezentációja állítható be, a kapott adatokon pedig további feldolgozási lépések végezhetők el, mint például megoszlásvizsgálatok, többszintű gyakorisági listák, többszavas kifejezések, kollokációk, igei argumentumok kinyerése.

6. Összegzés

Az MNSz. hivatkozási és látogatottsági adatai alapján egyértelmű, hogy az adatbázis a mai napig megkerülhetetlen forrása minden olyan kutatás-

[2] Inside, Outside, Beginning: szerkezet kezdő, szerkezeten belüli, szerkezeten kívüli elem.

nak és fejlesztésnek, amely magyar nyelvi adatot használ fel. A Szövegtár létrehozásával foglalkozó projekt hosszú időn keresztül a Korpusznyelvészetiből Nyelvtechnológiaivá vált osztály, de egyben a Nyelvtudományi Intézet zászlóshajója volt. Váradi Tamásnak az általa megalapított és irányított osztály központi tevékenységével kapcsolatos, a 90-es évek végén megfogalmazott jövőképe teljes mértékben beigazolódott.

Bibliográfia

Csendes, D., Csirik, J., Gyimóthy, T.: The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. In: Sojka, P., Pala, K., Kopecek, I. (szerk.) Text, Speech and Dialogue: 7th International Conference, TSD. pp. 41–47. Springer (2004)

Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I., Trón, V.: A Szószablya projekt. In: Alexin Z., Csendes D. (szerk.) Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem (2003)

Halácsy, P., Kornai, A., Oravecz, Cs.: HunPos – an open source trigram tagger. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague. (2007)

Halácsy, P., Kornai, A., Oravecz, Cs., Trón, V., Varga, D.: Using a morphological analyzer in high precision POS tagging of Hungarian. In: Proceedings of LREC 2006, pp. 2245–2248. (2006)

Kupietz, M.: Near-Duplicate Detection in the IDS Corpora of Written German. Technical Report IDS-KT-2006-01, Institut für Deutsche Sprache (2005)

Kučera, H., Francis, W. N.: Computational Analysis of Present-Day American English. Brown University Press, Providence, RI. (1967)

Miháczi, A., Németh, L., Rácz, M.: Magyar szövegek természetes nyelvi előfeldolgozása. In: Alexin Z., Csendes D. (szerk.) Magyar Számítógépes Nyelvészeti Konferencia. pp. 38–43. Szegedi Tudományegyetem (2003)

Oravecz, Cs., Dienes, P.: Efficient stochastic part of speech tagging for Hungarian. In: Rodríguez, M. G., Suarez Araujo, C. P. (eds.) Proceedings of the Third International Conference on Language Resources and Evaluation. pp. 710–717. ELRA, Las Palmas (2002)

Parker, R., Graff, D., Kong, J., Chen, K., Maeda, K.: English Gigaword Fifth Edition. Linguistic Data Consortium. (2011)

Pomikalek, J.: Removing Boilerplate and Duplicate Content from Web Corpora. Doktori disszertáció, Masaryk University, Faculty of Informatics, Brno. (2011)

Prószéky, G., Tihanyi, L.: Humor – A morphological system for corpus analysis. In: Rettig, H. (ed.) Proceedings of the first TELRI seminar in Tihany. pp. 49–158. Budapest (1996)

Trón, V., Gyepesi, Gy., Halácsy, P., Kornai, A., Németh, L., Varga, D.: Hunmorph: open source word analysis. In: Proceedings of the ACL 2005 Workshop on Software. pp. 77–85. The Association for Computational Linguistics (2005)

Váradi, T.: The Hungarian National Corpus. In: Rodríguez, M. G., Suarez Araujo, C. P. (eds.) Proceedings of the Third International Conference on Language Resources and Evaluation. pp. 385–389. ELRA, Las Palmas (2002)

Tamás Váradi and META-NET

Sabine Kirchmeier[1]

[1] European Federation of National Institutions for Language
sabine.kirchmeier@gmail.com

1. Introduction

In his longstanding career, Tamás Váradi has participated in numerous projects on the European scene, but probably the one with the greatest impact on the development of language technology in Europe is META-NET. META is an acronym for Multilingual Europe Technology Alliance, and the network aims at bringing together researchers, commercial technology providers, private and corporate language technology users, language professionals and other information society stakeholders. META-NET is a joint effort towards furthering language technologies in Europe.

2. The key components of META-NET

META-NET started as a network of excellence in 2010 pursuing the following goals:

1. fostering a dynamic and influential community around a shared vision and strategic research agenda (META-VISION),
2. creating an open distributed facility for the sharing and exchange of resources (META-SHARE),
3. building bridges to relevant neighbouring technology fields (META-RESEARCH)[1]

The network consisted of 60 members in 34 European countries, one of these the Research Institute for Linguistics at the Hungarian Academy of Sciences represented by Tamás Váradi.

2.1. META-VISION

META-VISION comprised two main activities: 1. addressing European decision makers through the development and promotion of a strategic

[1] http://www.meta-net.eu/

DOI: https://doi.org/10.18135/VT70.6

research agenda (Rehm and Uszkoreit, 2013), and 2. conducting a large and comprehensive study on 30 European languages and the level of support they receive through language technologies, the META-NET White Papers (Rehm and Uszkoreit, 2012). The two initiatives together turned out to provide an excellent basis for creating an understanding and gaining support among decision makers for the necessity of developing better language technology for all European languages.

2.1.1. The strategic research agenda

The strategic research agenda, published in 2013, set out to describe what kind of technological innovations could be expected by 2020 and what role language technology could play as part of these innovations. Robotics and AI were quite optimistically pointed out as fields expected to profit immensely from language technology:

> "Within this decade, specialised mobile robots will be deployed for personal services, rescue missions, household chores, and tasks of guarding and surveillance. Natural language is by far the best communication medium for natural human-robot interaction. By 2020 we will have robots around us that can communicate with us in human language, but their user friendliness and acceptance will largely depend on progress in LT research in the coming years" (Rehm and Uszkoreit, 2013: 35).

2.1.2 The META-NET White Paper Series

The META-NET White Papers (Rehm and Uszkoreit, 2012) describe the technological status of 30 European languages and the level of language technology support available for each of them. 200 experts from all over Europe, the META-NET Network of Excellence, contributed to this comprehensive study that discusses the threats and opportunities for the languages in question.

The key results and the cross-language comparison of the collected data sent a clear message that for most languages, except for English, it was extremely urgent that efforts were made to bring them up to a level where they could be preserved from digital extinction. The Hungarian language was among the languages in the danger zone with only fragmentary support for machine translation, speech processing and text analysis. For text and speech resources, Hungarian, together with Czech,

Dutch, French, German, Italian, Polish, Spanish and Swedish, was reported to have moderate support.

The analyses of the different languages served as an excellent starting point for public initiatives to create better language technology support for many languages.

2.2. META-SHARE

The collection and sharing of language resources is still today at the core of developing language technology. The demand for more and better text and speech resources has grown drastically, especially due to the use of AI-techniques during the last decade. The second META-NET initiative, META-SHARE, was established to address this problem by facilitating the sharing of resources, whereas the collection of resources was mainly the responsibility of national programmes.

2.3. META-RESEARCH

Finally, META-RESEARCH established several working groups, research workshops and a collection of online tutorials, mainly on machine learning for MT, but also other topics to support the development of language technology expertise for the languages involved.

3. META-NET and CESAR

Soon after its formation, META-NET succeeded in interlinking four EC-funded projects, resulting in a Europe-wide cooperation of computational linguistic and NLP communities. One of the projects involved was CESAR (Central and South-East European resources) (Váradi, 2013). It was funded by EC and national funding sources. The project started on 1st February 2011, its duration was 24 months, and the coordinator was Tamás Váradi.

The central objective of the project was to produce and make available a comprehensive set of language resources and tools covering Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak. CESAR was not only about the creation of new resources but also about enhancing existing resources and tools, for instance their size, coverage, accuracy, compliance with current standards for interoperability, and regarding licencing and IPR issues.

A huge effort was made to make the linguistic development environment NOOJ freely available on all platforms allowing linguists to formalize several levels of linguistic phenomena: typography and spelling; lexicons of simple words, multiword units and discontinuous expressions; inflectional, derivational and productive morphology; local and structural syntax, transformational and semantic analysis and generation.[2]

By the end of the project, the resources and tools developed by CESAR were made available through META-SHARE, thus contributing to the extension of the linguistic coverage of the platform and ensuring the availability of key resources for the development of improved language technology and AI applications for Central and South East European languages.

4. The impact of META-NET on European language technology

META-NET and its associated projects came to play an extremely important role for the development of language technology for European languages for several reasons. First, META-NET provided the basis for a strong European language technology community that worked together instead of competing with one another. Second, META-SHARE stimulated the collection and exchange of language resources for commercial use, in contrast to collections in other networks, such as CLARIN, which were more focussed on resources for research purposes. Finally, the META-NET White Paper Series gave the participating language communities a well-documented offset for the discussion about the future of languages and language technology all over Europe (Rehm et al., 2014).

4.1. A European LT community

The META-network extended constantly, not least through a long series of events and conferences (META-FORUM etc.)[3] and through the inclusion of new stakeholders, individual researchers, companies, and organisations such as the European Federation of National Institutions for Language (EFNIL) and the Network to Promote Linguistic Diversity (NPLD). In this way, META-NET was able to engage and include stakeholders in all European countries and to present itself to the European Commission as a strong language technology community with an

[2] http://www.nooj-association.org/
[3] http://www.meta-net.eu/events

impressive network of supporters consisting of private vendors, public institutions, and researchers all over Europe.

Over the years, the network continued to grow as its members participated in follow-up projects such as the EU-project CRACKER (2015–2017).[4] CRACKER's objectives were, among others, preparing and publishing research and innovation agendas (Rehm, 2015). It managed to establish the Cracking the Language Barrier federation,[5] a kind of umbrella initiative for European language technology projects and organisations. The cooperation also formed the basis of the European Language Grid.

Many of the META-NET members also participate in EU's European Language Resource Coordination project (ELRC) aiming at collecting language resources and providing a platform for sustainable language data sharing to support language equality in multilingual Europe, especially the Digital Single Market.[6]

The network created through META-NET also made it possible to conduct a survey (Rehm and Hegele, 2018), which covered more than 600 respondents from more than 50 countries working on language technology, emphasising the need of a programme specifically designed to develop the technology that could meet the linguistic challenges in Europe.

Many members of the network are involved in the latest European language technology project, European Language Equality (ELE), with the goal to establish a roadmap for the development of sustainable language technology for all European languages by 2030.[7]

4.2. Resources in META-SHARE

In 2012, there were 1248 – in 2020, there were 2888 language resources, tools, or services accessible through META-SHARE distributed over 100+ languages, four main resource types (corpus, lexical/conceptual model, tool/service, language description) and four main media types (text, audio, image, video). Currently, the most frequently viewed and downloaded datasets are those containing semantic annotations and gold standards. This clearly indicates the direction that language technology is taking towards becoming an integrated part of the development of artificial intelligence applications.

[4] http://www.cracker-project.eu
[5] http://www.cracking-the-language-barrier.eu
[6] https://lr-coordination.eu
[7] EU-CALL: Developing a strategic research, innovation and implementation agenda and a roadmap for achieving full digital language equality in Europe by 2030 (PPPA-LANGEQ-2020)

4.3. Political attention

The cross-comparison of the digital fitness of the participating languages regarding text analysis, speech processing, MT and language resources in the META-NET white papers clearly made an impression on the government in many countries and has led to the development of strategies and plans for the advancement of high-quality language technology, for instance in Denmark, Finland, Iceland, Latvia, Norway and Sweden.

In the EU, it led to a hearing initiated by the Scientific Foresight Unit (STOA) of the European Parliament in 2017 (STOA, 2017), who also commissioned the study Language Equality in the digital age – Towards a Human Language Project, published in March 2017 and presented to the European Parliament on 11 Sept. 2018.[8] The EP adopted the report, with an overwhelming majority of 592 votes in favour, 45 against, and 44 abstentions.

The newly initiated ELE Project (2021–2022) is the first step towards implementing the vision of language equality laid out in the STOA-report. It is envisaged to be a multidisciplinary initiative including stakeholders from research institutions, industry, the public sector and civil society, collaborating on European, national and regional level. The primary goal is the preparation of the European Language Equality Programme, specified in the form of a strategic research, innovation and implementation agenda and a roadmap for achieving full digital language equality in Europe by 2030.

With his long-standing experience from not least META-NET and CESAR and his strong and untiring dedication to language technology, Tamás Váradi is of course also involved in the ELE project laying out the path for European language technology in the future. In fact, he is involved both in his capacity as head of his institute and as general secretary of EFNIL as well. There is no doubt that his dedicated engagement in language technology through the years is of immense importance for the digital future of the Hungarian language.

References

[8] Language equality in the digital age (A8-0228/2018, P8_TA-PROV(2018)0332)

Rehm, G., Uszkoreit, H. (eds.) META-NET White Paper Series: Europe's Languages in the Digital Age. Springer, Heidelberg/New York/Dordrecht/London (2012) www.meta-net.eu/whitepapers. [31 volumes on 30 European languages.]

Rehm, G., Uszkoreit, H. (eds.) META-NET Strategic Research Agenda for Multilingual Europe 2020. Presented by the META Technology Council. Springer, Heidelberg, New York etc. (2013) http://www.meta-net.eu/sra

Rehm, G., Uszkoreit, H., Dagan, I., Goetcherian, V., Dogan, M. U., Mermer, C., Váradi, T., Kirchmeier-Andersen, S., Stickel, G., Jones, M. P., Oeter, S., and Gramstad, S.: An Update and Extension of the META-NET Study "Europe's Languages in the Digital Age". In Laurette Pretorius, et al. (eds.) Proceedings of the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (CCURL 2014), pp. 30–37. Reykjavik, Iceland (2014)

Rehm, G.: Cracking the Language Barrier for a Multilingual Europe. In: Nuolijärvi, P., Stickel, G. (eds.) Language Use in Public Administration – Theory and practice in the European states. Contributions to the Annual Conference 2015 of EFNIL in Helsinki. pp. 41–58. European Federation of National Institutions for Language, Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary (2015)

Rehm, G., Hegele, S.: Language Technology for Multilingual Europe: An Analysis of a Large-Scale Survey regarding Challenges, Demands, Gaps and Needs. In: Calzolari, N. et al. (eds) Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018), Miyazaki, Japan. pp. 3282–3289 (2018)

Váradi, T.: Veni, Vidi, Vici: The Language Technology Infrastructure Landscape after CESAR. In: Gajdošová, K., Žáková, A. (eds.) Natural Language Processing, Corpus Linguistics, E-learning. Seventh International Conference Bratislava, Slovakia, 13–15 November 2013 Proceedings, pp. 261–279 RAM-Verlag, Lüdenscheid (2013)

Az EFNILEX első szakasza

Héja Enikő[1]

[1] Nyelvtudományi Intézet
heja.eniko@nytud.hu

1. Bevezetés

A cikkben ismertetett munka az EFNIL által finanszírozott EFNILEX projekt első szakasza, amely 2008-tól 2012-ig tartott. A projekt azt vizsgálta, hogy a nyelvtechnológiai módszerek és eszközök – különös tekintettel a párhuzamos korpuszokon végzett szóillesztésre – mennyiben képesek támogatni a szótárkészítési folyamatot. A szótárkészítés automatikus támogatása elsősorban a kevéssé használt nyelvek esetében bír jelentőséggel, hiszen az ilyen nyelvpárokra íródott szótárakra alacsony a kereslet, így az ilyen munkálatok finanszírozása is korlátozott. A bemutatandó munka eredeti célja egy közepes méretű (kb. 15 000 szócikk) litván–magyar szótár létrehozása volt. A munkafolyamat részeként tesztelési célokra a magyar–szlovén nyelvpárt is vizsgáltuk.

A projektben részt vettek: František Čermak (Institute of the Czech National Corpus, Charles University), John Simpson (Oxford English Dictionary) és Jolanta Zabartskaite (Institute of the Lithuanian Language). A projektet Váradi Tamás koordinálta az MTA Nyelvtudományi Intézet részéről.

Az elvégzett munkából egy disszertáció is született Váradi Tamás témavezetésével (Héja, 2016), melynek főbb eredményeit az *International Journal of Lexicography* is közölte (Héja, 2017).

Az EFNILEX projekt célja egy megfelelő lefedettségű és pontosságú lexikai erőforrás előállítása volt, amely emberi utószerkesztési munkálatokat is igényel. Vagyis célunk az volt, hogy a lexikográfusok számára olyan erőforrásokat biztosítsunk, amelyek a lehető legjobban csökkentik a teljes értékű, emberi felhasználásra alkalmas szótárak elkészítéséhez szükséges munkát. A fenti elvárásoknak megfelelő automatikusan generált erőforrásokat protoszótáraknak fogjuk nevezni a cikk hátralevő részében. Az automatikusan létrehozott protoszótárak az efnilex.efnil.hu weboldalon kérdezhetőek le.

DOI: https://doi.org/10.18135/VT70.7

Az általunk javasolt módszer alapját párhuzamos korpuszokon végzett automatikus szóillesztés képezi. Bár az automatikus szóillesztést széles körben használták szótárfejlesztésre elsősorban a gépi fordítás területén, amennyire tudjuk, a kutatás előtt ezt a megközelítést nem használták emberi felhasználásra szánt szótárak készítésének támogatására lexikográfiai projektekben.

A megfelelő módszer kiválasztásakor egyik fő szempontunk az volt, hogy a lehető legnagyobb mértékben csökkentsük a lexikográfusi nyelvi intuíció szerepét a szótárkészítési folyamat során, ezért a protoszótárakat felügyelet nélküli tanulási algoritmussal akartuk létrehozni. Elsősorban ezért választottuk ezt a módszert például a *hub-and-spoke* (Martin, 2007) modellel szemben, amelynek lényege, hogy már létező egynyelvű adatbázisokat (pl. egynyelvű értelmező szótárakat vagy wordneteket) köt össze egy olyan sok erőforrással rendelkező közbülső nyelv felhasználásával, amely rendelkezik kétnyelvű szótárakkal mind a forrásnyelv mind a célnyelv tekintetében.

A cikk hátralévő részében a munkafolyamatot, valamint az elért eredményeket mutatom be.

2. A munkafolyamat leírása

A munkafolyamtot számos cikkben ismertettem már (pl. Héja, 2010, Héja és Takács, 2012), így ezt most csak vázlatosan fejtem ki.

A protoszótárak készítése minden vizsgált nyelvpár esetében három lépésből állt. Az első szakaszban elkészítettük a párhuzamos korpuszok egységes morfológiai annotációval ellátott XML-verzióját a vizsgált nyelvpárokra. A második szakaszban a párhuzamos korpuszok lemmatizált változatából szóillesztéssel létrehoztuk a protoszótárakat. A harmadik szakaszban kiértékeltük a protoszótárakat.

2.1. A párhuzamos korpuszok elkészítése

A projekt során az eredetileg célul kitűzött nyelvpárok mellett (litván–magyar, szlovén–magyar) további nyelvpárokat, a francia–holland, illetve az angol–magyar nyelvpárokat is vizsgáltunk. Mivel ez utóbbiak esetében volt elérhető párhuzamos korpusz (DPC [Macken és mtsai., 2007], Hunglish [Varga és mtsai., 2005]), csak a litván–magyar, illetve szlovén–magyar nyelvpárokra építettünk párhuzamos korpuszt. Az első szakasz során a szövegeket összegyűjtöttük, normalizáltuk és morfológiailag elemeztük. A párhuzamosítást a hunalign

(Varga és mtsai., 2005) mondatillesztővel végeztük el. A párhuzamosítást egynyelvű szövegek lemmatizált változatain végeztük el. Következő lépésben elkészítettük a morfológiailag elemzett párhuzamos korpuszok egységes XML-reprezentációját. A munkaszakaszhoz kapcsolódó legfontosabb tapasztalat az volt, hogy a javasolt módszer legnagyobb nehézségét a kevéssé használt nyelvpárok esetében a digitálisan elérhető párhuzamosítható szövegek mennyisége jelenti. Amennyiben a megfelelő méretű párhuzamos korpusz rendelkezésre áll, a protoszótár már könnyen előállítható.

2.2. A szóillesztés

A második szakaszban a protoszótárakat állítottuk elő a párhuzamos korpuszok alapján. A szótárkinyerésre a GIZA++-t (Och és Ney, 2003) választottuk a számos versengő módszer közül (pl. Ribeiro és mtsai., 2000). A módszer lényege, hogy a szóillesztés elvégzése során a forrásnyelv és a célnyelv lemmái között feltételes valószínűséget becsül ($P(lemma_{cél}|lemma_{forrás})$). Az így becsült szópárok közötti feltételes valószínűségek képezik a protoszótárak alapját. A lehetséges szótárkinyerő algoritmusok közül azért éppen ezt választottuk, mert azt gondoltuk, hogy aszimmetrikus jellegénél fogva a feltételes valószínűségek becslése különösen alkalmas megfordítható, kódoló szótárak előállítására. Erről bővebben lesz még szó jelen írás 3.2 fejezetében.

2.3. A kiértékelés

A munkafolyamat harmadik szakasza a kiértékelés volt. A kiértékelést több lépésben végeztük. Az első lépésben az annotátorok közötti egyeztetés után megállapítottuk, hogy milyen típusú hibák fordulnak elő a fordítási jelöltek között. Ezen tapasztalatok alapján készítettünk egy kiértékelési útmutatót is, amely során törekedtünk arra, hogy disztribúciós alapokon határozzuk meg az egyes hibatípusokat. Ezzel az volt a célunk, hogy a lehető legjobban lecsökkentsük a szubjektív egyéni nyelvi értékítélet szerepét a kiértékelés során.

A kiértékelés második lépése már az útmutató alapján zajlott. Ennél a lépésnél már nemcsak a fordítási valószínűséget vettük figyelembe, de a forrásnyelvi és célnyelvi lemmák előfordulási gyakoriságát is felvettük paraméternek.

A kiértékelés második szakasza alapján a következő főbb következtetéseket tettük: (1) Szükséges a lemmáknak egy minimális előfordulási

gyakorisága ahhoz, hogy becsülhető legyen a valószínűség. (2) Ha van megfelelő mennyiségű adat, akkor általában igaz az, hogy minél nagyobb a fordítási valószínűség, annál jobb a fordítás. (3) De magas fordítási valószínűség esetén is lehet nagy a hibás fordítások aránya: gyakran előforduló forrásnyelvi lemma és ritka célnyelvi fordításjelölt esetén, ha a forrásnyelvi és célnyelvi lemmák sokszor fordulnak elő együtt párhuzamos mondatokban. Azért, hogy az ilyen eseteket kiszűrjük, az eddigi paraméterek mellett figyelembe vettük még a forrásnyelvi és célnyelvi lemmák gyakoriságának hányadosát is: ennek egy előre meghatározott küszöbérték alatt kellett maradnia. (4) A következő megfigyelésünk az volt, hogy a forrásnyelvi lemmák gyakoriságai és a fordítási valószínűségek „fordítottan arányosak": azaz minél gyakrabban fordul elő a forrásnyelvi lemma, annál kisebb fordítási valószínűségek is még jó fordításokat eredményeznek.

Így harmadik lépésben egy sávos kiértékelést is elvégeztünk, amely során a forrásnyelvi lemma növekvő gyakorisági intervallumaihoz csökkenő valószínűségi küszöbértékeket rendeltünk. Azt találtuk, hogy a fordítási párok ilyen szűrése alkalmas a fedés növelésére is.

3. A projekt eredményei

3.1. Gyakorlati eredmények

A projekt során lekérdezhető protoszótárakat hoztunk létre négy nyelvpárra: magyar–litván (v.v.), magyar–szlovén (v.v.), francia–holland (v.v.), angol–magyar (v.v.). A lekérdezhető protoszótárak megalkotása során az adatbázisok mellett kialakítottunk egy – a hagyományostól némileg eltérő – lekérdezőfelületet is, amely lehetővé teszi a választott módszer adatvezérelt jellegéből fakadó új információk megjelenítését, illetve lekérdezését.

A protoszótárak számos sajátossággal bírnak. Először is: a választott módszer miatt megfordíthatók, így nem négy, hanem nyolc protoszótárat hoztunk létre. A protoszótárak az efnilex.efnil.org weboldalon kérdezhetők le.

A protoszótárak kódoló szótárak, így különösen alkalmasak arra, hogy szövegek írásában segítsék a felhasználót azáltal, hogy hasznos információkat nyújtanak a fordítás helyes használatára vonatkozóan. Egyfelől megjelenítik azokat a párhuzamos kontextusokat, amelyben a forrásnyelvi és a célnyelvi szavak előfordulnak. Ezen túl a protoszótárak segítik a fordítás helyes használatát azzal is, hogy az előfordulási gyakorisá-

gok alapján megbecsülik, hogy a fordítási jelölt használati köre szűkebb-e vagy tágabb, mint a forrásnyelvi szóé. Előbbi esetben a szöveg megalkotásakor a célszó kontextusaira kiemelt figyelmet kell fordítani. A lekérdezhető protoszótárak további érdekessége, hogy testre szabhatók annak függvényében, hogy milyen felhasználói csoportot céloznak meg. Ha csak a leggyakoribb szavak fordításait kérdezzük le magas feltételes valószínűséggel, akkor megkapjuk egy nyelv alapszókincsét kevés, ám biztosan jó fordítási jelölttel. Ez a beállítás kezdő nyelvtanulók számára ajánlott. Ezzel szemben a protoszótárakat úgy is testre szabhatjuk, hogy a ritkább szavak nem tipikus fordításait is megjelenítsék. Ebben az esetben több lesz a hibás fordítási jelölt, de mivel ezekre a fordításokra már elsősorban a biztos nyelvismerettel rendelkezők kíváncsiak, ők kézzel kiszűrhetik a helytelen fordítási jelölteket.

A projekt gyakorlati eredményei közé soroljuk az egységes morfológiai annotációval ellátott litván–magyar, szlovén–magyar és angol–magyar párhuzamos XML-korpuszokat is. A párhuzamos korpuszok méretét az 1. táblázatban adjuk meg:

1. táblázat. A morfológiailag annotált párhuzamos XML-korpuszok mérete.

	Magyar	**Litván**	**Magyar**	**Szlovén**	**Magyar**	**Angol**
Token	4.813.956	4.141.521	723.857	809.448	6.921.127	8.312.795
Mondat	319.489	320.678	40.926	42.659	494.044	494.044
Fordítási egység	304.419		38.791		494.044	

3.2. Elméleti eredmények

A projekt legfontosabb elméleti eredménye, hogy a javasolt módszer, vagyis a fordítási párok automatikus tanulása párhuzamos korpuszokból feltételes valószínűségek becslésével, számos előnnyel rendelkezik a hagyományos és korpuszalapú kétnyelvű lexikográfiai módszerekkel szemben is. Ezek közül a legfontosabb, hogy a javasolt módszer a forrásnyelvi oldalon kiküszöböli a lemmákhoz tartozó egyes jelentések elkülönítésének problémáját. Továbbá, a módszer lehetővé teszi a fordítási reláció korpusz adatokon való kvantifikálható újraértelmezését. A szakirodalom (pl. Atkins és Rundall, 2010, Adamska-Sałaciak, 2010) alapján azt találtuk, hogy a fordítási reláció általában valamilyen értelemben aszimmetrikus és fokozatos. Azt állítjuk, hogy a hagyományos relációs

felfogás helyett a fordítási relációra érdemes feltételes valószínűségként gondolni. Hiszen a feltételes valószínűség megragadja a fordítási reláció aszimmetrikus és fokozatos jellegét. Sőt ezen túlmenően ez a matematikai konstrukció számot ad arról a speciális esetről is, amikor a fordítási reláció szimmetrikus. Ez a tökéletes fordítási ekvivalencia esetében áll fenn.

3. Összefoglalás

A cikkben az EFNILEX projekt első szakaszának (2008–2012) főbb eredményeit ismertettem, melyet az EFNIL tagszervezeteként végeztünk Váradi Tamás koordinálásával. Számos magyar és nemzetközi publikáció mellett a disszertációm is ebből a munkából született, melynek témavezetője szintén Váradi Tamás volt. A disszertáció főbb elméleti eredményei az *International Journal of Lexicography*-ban is megjelentek.

Végezetül néhány személyes gondolatot szeretnék leírni. Az egyetemről frissen kikerülve sokunknak volt a Korpusznyelvészeti, majd később a Nyelvtechnológiai Osztály az első munkahelye. Vezetési stílusából fakadóan Tamás gyakran előlegezett bizalmat nekünk a feladatok kiosztása során. Bár ennek kapcsán olykor előfordult velem, hogy azt éreztem, túl nagy ez a kabát, egyúttal ez nagyon motiváló is volt. Remélem, hogy Tamás is úgy gondolja, hogy ehhez a megelőlegezett bizalomhoz a legtöbb esetben sikerült felnőnünk.

A projekteket és a kapcsolódó kutatásokat gyakran mutathattuk be neves külföldi konferenciákon, amelyet az ünnepelt mindig támogatott anyagilag is, ennek köszönhetően már pályánk elején bekapcsolódhattunk a nemzetközi vérkeringésbe. Így sokunkat Tamás indított el a nyelvtechnológiai pályán. Ezért nagyon hálás vagyok, és ezzel a rövid írással szeretnék boldog 70. születésnapot kívánni Neki. Kedves Tamás, Isten éltessen!

Bibliográfia

Atkins, B. T. S., Rundell, M.: *The Oxford Guide to Practical Lexicography*. Oxford University Press, Oxford (2008)

Adamska-Sałaciak, A.: Examining Equivalence. *International Journal of Lexicography* 23/4, 387–409 (2010)

Héja E.: Dictionary Building based on Parallel Corpora and Word Alignment. In: Dykstra, A. and Schoonheim, T. (eds) Proceedings of the XIV. EURALEX International Congress. pp. 341–352. Fryske Akademy, Afûk, Ljouwert (2010)

Héja, E.: The Usability of Language Technology Methods and Parallel Corpora in Bilingual Lexicography. Quantifying Translational Equivalence. PhD-értekezés (2016)
Héja, E.: Revisiting Translational Equivalence: Contributions from Data-Driven Bilingual Lexicography International Journal of Lexicography 30/4, 483–503 (2017)
Héja, E., Takács, D.: Automatically Generated Customizable Online Dictionaries. In: Daelemans W. et al. (eds.) Proceedings of EACL2012. pp. 51–57. The Association for Computer Linguistics (2012)
Macken, L., Trushkina, J., Paulussen, H., Rura, L., Desmet, P., Vandeweghe, W.: Dutch Parallel Corpus. A multilingual annotated corpus. In: Davies, M., Rayson, P., Hunston, S., Danielsson, P. (eds.) Proceedings of Corpus Linguistics 2007. University of Birmingham, Birmingham, United Kingdom (2007)
Martin, W.: Government Policy and the Planning and Production of Bilingual Dictionaies: The 'Dutch' Approach as a Case in Point, International Journal of Lexicography 20/3, 221–237 (2007)
Ribeiro, A., Pereira Lopes, G., Mexia, J.: Extracting Equivalents from Aligned Parallel Texts: Comparison of Measures of Similarity. In: Monard M.C., Sichman J. S. (eds.) Advances in Artificial Intelligence. IBERAMIA 2000, SBIA 2000. Lecture Notes in Computer Science, vol 1952. pp. 339–349. Springer, Berlin, Heidelberg (2000)
Och, F. J.; Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics 29/1, 19–51 (2003)
Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., Nagy, V.: Parallel corpora for medium density languages. In: Angelova, G., Bontcheva, K., Mitkov, R. Nicolov, N., Nikolov, N. (eds.) Proceedings of the RANLP 2005. pp. 590–596. Borovets, Bulgaria (2005)

Serving Multilingual Europe

Svetla Koeva[1]

[1] Institute for Bulgarian Language "Prof. L. Andreychin", Bulgarian Academy of Sciences
svetla@dcl.bas.bg

1. Introduction

Serving Multilingual Europe is the most precise wording that characterizes the project CESAR (**CE**ntral and **S**outh-East Europe**A**n **R**esources), which was formulated by prof. Tamás Váradi, the coordinator of the project. His paper entitled *Serving Multilingual Europe: The CESAR Project* is the most prominent study included in the book *Language Resources and Technologies for Bulgarian*, published by "Professor Marin Drinov" Publishing House of the Bulgarian Academy of Sciences (Váradi, 2014) in 2014 and dedicated to the results of two major European projects, one of which is the project CESAR. The CESAR project (Central and South-East European Resources)[1] was funded by the European Commission through the ICT Policy Support Programme, Grant agreement no.: 271022. The runtime of the project was from February 1st, 2011 until January 31st, 2013.

At the introduction of his paper, Prof. Váradi presented in depth the rationale motivating the project and its origins, outlining the issue of tackling language as one of the most prominent challenges in the age of digital communication, which is becoming increasingly multimodal and multilingual. Prof. Váradi stressed "the mission of language technologies to ensure that people can communicate with each other as well as with automated services via digital devices in the most natural, unconstrained manner possible" (Váradi, 2014: 9). In fact, the author predicted the extreme need for the collection and processing of big, multimodal and multilingual data, a need that still exists as language technology crucially depends on data, and postulated the requirement for language resources that are applicable, useful and easily available for language technologies. All this motivated the building of a language technology infrastructure in a pan-European coordinated manner, the META-NET (A Network of

[1] http://cesar.nytud.hu

DOI: https://doi.org/10.18135/VT70.8

Excellence forging the Multilingual Europe Technology Alliance: META). The consortium consisted of nine partners covering six languages, with five of them belonging to the Slavic group of languages (Bulgarian, Croatian, Polish, Serbian, Slovak), and the sixth, Hungarian – a Finno-Ugric language. The CESAR project, as part of META-NET, intended to address the issue of the sporadic development of language resources and tools for the so-called less-resourced languages by enhancing, upgrading, standardising and cross-linking a wide variety of language resources and tools and making them available by means of an open infrastructure.

2. META-SHARE and CESAR project

The central objective of the CESAR project was "to produce and make available a comprehensive set of language resources and tools covering Bulgarian, Croatian, Hungarian, Polish, Serbian, and Slovak" (Váradi, 2014: 11). The focus was on diverse types of resources and tools: mono- and multilingual speech databases, mono- and multilingual corpora, dictionaries, wordnets, natural language processing tools: tokenisers, lemmatisers, taggers, parsers, named-entity recognizers and so on. The activities were directed not so much to the creation of new language resources and language technology tools, but rather to "the enhancement of existing resources and tools (in size, coverage, precision, recall, accuracy), the adaptation of resources and tools to become compliant with the agreed standards for interoperability, as well as the upgrade of resources and tools by combining them with other resources and tools" (Váradi, 2014: 12). Here the role of professor Váradi in the preparation of the project's proposal has to be emphasized. Thanks to his skills to coordinate the activities of people with different skills, experience and knowledge; to analyze the objective conditions and to offer the most appropriate solutions in relation to the needs, potentials and expectations; to formulate clearly and precisely the project objectives, work packages, and expected results the successful realization of the project was presupposed and to a great extent predetermined.

Within the CESAR project the META-SHARE platform[2] has been established and populated with language resources and tools (in collaboration with other members of META-NET network of excellence). The META-SHARE platform organises an open network of repositories for

[2] http://www.meta-share.org

sharing language data, tools and web services. As META-SHARE is implemented in the framework of the META-NET Network of Excellence, the CESAR project contributed to the building of META-SHARE and its population with language resources, language processing and annotation tools and technologies, and services. Prof. Váradi explained in his paper that servers linked to META-SHARE form a chain of nodes, organized into a hierarchical structure: managing nodes are synchronized, and provide the META-SHARE metadata and resources, while network nodes are not synchronized, but harvested by a managing node. Within the project, for each language included in the project at least one network node was established and in 2021 the META-SHARE platform itself and many of CESAR network nodes are operating ensuring long-term sustainability (for example, the Hungarian and the Bulgarian nodes). The CESAR project provided through META-SHARE 251 language resources and language processing tools for Bulgarian, Croatian, English, Hungarian, Polish, Serbian, Slovak, and other languages: 66 tools and services; 120 mono- and multilingual corpora, and 65 lexical and conceptual resources. Prof. Váradi stressed within his paper that language resources created, improved and expanded within the Croatian project are at a high technological level, which allows for their direct integration in various applications based on language technologies, both for research and commercial purposes.

To provide the most suitable resources and tools, the project partners have developed a special methodology for the selection of resources. The consortium developed a list of four general indicators that were considered representative and indicative of the selection of language resources. The indicators determine the general requirements to which the selection should be subjected. Different sets of specific criteria have been defined for each indicator. We want to emphasize again the leading role of Prof. Váradi during the discussions and in the process of the selection of the most appropriate and balanced indicators. The general indicators were, for example: a) For upgraded resources: all selected resources are state-of-the-art specimens of their type for a given language; equally valuable representatives are all included in the selection; etc.; b) For extended/linked resources: the extension of resources provides considerable value to the community, at least on regional level; the emphasis is on providing building blocks to the existing tools rather than major restructuring; etc.; c) For resources aligned across languages: no more than one tool of a certain type for each language is used; whenever applicable,

the largest set of languages is selected; etc. The general indicators were combined with the so-called Total Point Value (Maegaard, 2004) concerning the availability, quality, quantity and standards for language resource and tools under selection. Also the IPR principles and legal issues were taken into consideration, promoting the use of open data and following the Creative Commons and Open Data Commons principles.

Prof. Váradi devoted a special place in his paper to the Bulgarian contribution to CESAR project which he classified as "vital to the success of the CESAR project" (Váradi, 2014: 17). He gave the pride of place to the Bulgarian National Corpus,[3] which is "not only impressive but stands unique within the CESAR languages in terms of size and composition" (Váradi, 2014: 17). Prof. Váradi also recognised the Bulgarian WordNet; the Bulgarian Sense–annotated Corpus, Bulgarian-X Language Parallel Corpus, etc. He stressed that the Bulgarian contribution was not confined to resources unique for the Bulgarian language, but included resources made cross-lingual with the participation of all partners (Váradi, 2014: 18). It must be said that the special attention to the Bulgarian resources and tools is paid by Prof. Váradi due to the fact his paper is the key paper in a volume devoted to Bulgarian resources and tools produced within the CESAR project. Prof. Váradi has always paid equal attention to the project partners and has succeeded in an extremely delicate and appropriate way in both praising for success and in insisting on overcoming shortcomings, if any.

Prof. Váradi also described the activities of the partners in order to strengthen the position of language technologies by means of presenting the series of Language White Papers (Rehm and Uszkoreit, 2013), which describe the state-of-the-art overview of 30 European languages from the perspective of the maturity of language technologies.[4] For each of the project languages a special volume of the series was developed, where the White Papers shed light on the language technologies in Bulgarian, Croatian, Hungarian, Polish, Serbian, and Slovak in 2013, analysing criteria such as quantity, availability, quality, coverage, maturity, sustainability, and adaptability of available language resources and tools. Quantity was measured on the answers to the question: Does a tool/resource exist for the language at hand? And the more tools/resources existed, the higher the rating was. Availability was ranked by answering the question: Are tools/resources accessible, freely usable on any platform

[3] http://search.dcl.bas.bg
[4] http://www.meta-net.eu/whitepapers/overview

or only available at a high price or under very restricted conditions? Quality was assessed by answering the question: How well are the respective performance criteria of tools and quality indicators of resources met by the best available tools, applications or resources? Coverage was correlated with the answers to the questions: To what degree do the best tools meet the respective coverage criteria?; To what degree are resources representative of the target language or sublanguages? Maturity was measured with the answers to the questions: Can the tool/resource be considered mature, stable, ready for the market?; Can the best available tools/resources be used out-of-the-box or do they have to be adapted? Sustainability was related with how well the tool/resource could be maintained/integrated into current information technology systems and adaptability – with how well the best tools or resources could be adapted/extended to new tasks/domains/genres/text types/use cases, etc. and played a major role during and after the project in assessing the state of language resources for European readers. Even eight years later, the Language White Papers are important in their methodology, in the way they conduct the investigation and analyze the results, as well as in measuring the progress in creating language resources and tools for individual languages. No language was considered to have "excellent support", only English was assessed as having "good support", followed by languages such as Dutch, French, German, Italian and Spanish with "moderate support". Languages such as Bulgarian, Hungarian and Polish exhibited "fragmentary support" in 2012. As Prof. Váradi described in his paper, the Language White Papers "were used extensively by the project partners to disseminate information about META-NET and CESAR at the national level to different stakeholders, primarily through the set of CESAR roadshows, one-day high-level events each dedicated to one language" (Váradi, 2014: 19). These events represented an ideal opportunity to spread the Language White Papers as widely as possible: target users were representatives of research centres, small and large technology corporations, translation services and other users or producers of Language Technology, language communities and societies, and policy makers responsible for supporting research and innovation, economy and Information and Communication Technology. For example, the participants at the roadshow in Sofia in 2012 were over 80.

Prof. Váradi compared the available resources and tools for Bulgarian, Croatian, Hungarian, Polish, Serbian, Slovak before and after the end of the CESAR project: at the beginning of the project the initial list of tools

and resources potentially available for enhancement numbered 130 in total (Váradi, 2014: 22). According to the analysis made by Prof. Váradi, the most valuable progress was achieved in the field of multilingual and national corpora. Also, a special effort was made to render language independent resources and tools and to support cross-linguality both in the field of resources and of tools. Prof. Váradi emphasised that "progress was not only quantitative (extension, upgrade, new resources and tools), but also qualitative" (by means of intellectual property right clearance and carefully prepared and detailed resource metadata standardisation and development) (Váradi, 2014: 21). He described the number of requirements which were set up in order to meet the long-term sustainability of language resources and tools: careful selection of language resources by means of especially designed methodology; performing particular actions to ensure quality and quantity of the selected resources – upgrading, extending and linking the resources across languages; making resources visible and accessible through the META-SHARE platform and extensive metadata descriptions based on established standards.

3. The project leader

As Prof. Váradi concluded, the main role of the CESAR project was Serving Multilingual Europe. We can further generalise that the role of Prof. Váradi as project coordinator and key researcher in the successive European projects (*Multilingual Resources for CEF.AT in the legal domain* – MARCELL,[5] funded by Connecting Europe Facility / Telecommunications sector, 01.10.2018–31.03.2021 and *Curated Multilingual Language Resources for CEF AT Action* – CURLICAT[6] funded by Connecting Europe Facility / Telecommunications sector, 01.06.2020 – 31.05.2023) is invaluable. His endless and highly significant dedication to contributing to the development of powerful multilingual, cross-lingual and monolingual technologies for European languages; to facilitating the strengthening of the European language technology community, uniting industry, innovation and research; to being one of the most prominent language technology policy makers and visionaries across Europe and beyond have left an everlasting mark on the present and future of multilingual, technologically advanced Europe.

[5] https://marcell-project.eu
[6] https://curlicat.eu

References

Maegaard, B.: The NEMLAR Project on Arabic Language Resources. In: 9th EAMT Workshop, “Broadening horizons of machine translation and its applications”, 26–27 April 2004. Malta. pp. 124–128. European Association for Machine Translation (2004)

Rehm, G., Uszkoreit, H. (eds). META-NET Strategic Research Agenda for Multilingual Europe 2020. Springer, Heidelberg, New York etc. (2013)

Váradi, T.: Serving Multilingual Europe: The CESAR Project. In: Koeva. Sv. (ed.) Language Resources and Technologies for Bulgarian. pp. 9–28. “Professor Marin Drinov” Publishing House of the Bulgarian Academy of Sciences, Sofia (2014)

A nyelvi veszélyeztetettségről közérthetően: az INNET projekt

Bakró-Nagy Marianne[1], Oszkó Beatrix[1], Sipos Mária[1], Várnai Zsuzsa[1]

[1] Nyelvtudományi Intézet
{bakro, oszko.beatrix, sipos.maria, varnai.zsuzsa}@nytud.hu

1. Bevezetés

Tapasztalt kutatók tudják, hogy valamely kutatási témára rátalálni igen sokféleképpen lehet, kezdve a véletlen (vagy annak tűnő) felismeréstől egészen a szakirodalom állandó böngészése közben felismert problémáig. Vannak, akiket a határidők ihletnek meg és vannak, akikben egy pályázati kiírás megpillantása szüli a gondolatot. De olyan is akad, amikor az embert váratlanul és minden előzmény nélkül felhívja egy kollégája, mondván, van itt az Európa Tanácsnak ez a 7-es keretprogramja, nem akarjuk ezt megtölteni valamilyen finnugor tartalommal? A válasz sokféle lehet, de egy évtizede ez volt: de igen, akarjuk, megbeszéljük! Persze ilyen telefonhívás csak olyasvalakitől érkezhet, akinek megvan a sok évtizedes tapasztalata meg az éles szeme is ahhoz, hogy felismerje, infrastrukturális kutatások felépítéséhez és megvalósításához hogyan vezet az út. Esetünkben az INNET-hez, ahhoz a projekthez, amit ez a dolgozat, ha vázlatosan is, de bemutat. Mert a telefonhívást, amely természetesen Váradi Tamástól érkezett, hosszas megbeszélések, viták itthon és külföldön, nemzetközi nyári egyetemi órák és unalmas kéziratolvasások, mérgelődések (amit magunk között csak „ordítozás”-nak neveztünk), konzultációk a tanárokkal, nagy sikerű előadások és végül megkönnyebbülés követte – ahogyan az lenni szokott. És sok tanulás: amit mi megtanultunk arról, hogyan kell ezt voltaképpen csinálni, meg amit mi el tudtunk magyarázni a nyelvtechnológusoknak arról, mit is művelünk mi valójában. Nem ez volt az első és egyetlen közös munkánk, hiszen az Uralonet is a nyelvtechnológusokkal együtt készült, de az INNET honlapjának jelentősége – és ezt akkor talán még csak sejtettük, ám olyan biztosak nem lehettünk benne, mint ma vagyunk – az évek során egyre

DOI: https://doi.org/10.18135/VT70.9

csak nőtt, növekszik: érvényes ismereteket kínál egy erre kevéssé érzékeny és kevéssé elfogadó világban. Szóval köszönjük a hívást!

Az INNET projekt – Innovative Networking in Infrastructure for Endangered Languages – az Európa Tanács támogatásával, a 7-es Keretprogram 284415-as számú megállapodása eredményeképpen készült 2010 és 2014 között. A konzorcium tagjai voltak: Hollandia (Max Planck Institute), Lengyelország (Adam Miczkiewicz University), Németország (Universität zu Köln), Magyarország. Az MTA Nyelvtudományi Intézetében a projektvezetők: Bakró-Nagy Marianne – Finnugor és Nyelvtörténeti Osztály, Váradi Tamás – Nyelvtechnológiai és Alkalmazott Nyelvészeti Osztály. A magyar csapat tagjai: Duray Zsuzsa, Oszkó Beatrix, Sipos Mária, Szeverényi Sándor, Várnai Zsuzsa. A magyar honlaphoz kapcsolódó programozási feladatokat Herczog Zoltán végezte el. Az alábbiakban a projekt keretében megvalósult honlapot mutatjuk be, amelynek magyar változata a két osztály együttműködésében jött létre.

A honlap céljai a következők:

- Megismertetni a diákokkal a nyelvi veszélyeztetettség fogalmát, felkelteni érdeklődésüket a nyelvi diverzitás iránt, valamint felhívni figyelmüket a nyelvi dokumentáció fontosságára.
- Ezekről a témákról ismeretekkel ellátni a tanárokat és a diákokat, valamint megmutatni, hol juthatnak további információhoz.
- Segítséget nyújtani a tanároknak abban, hogy ezeket a témákat integrálhassák óráik anyagába.

A fejlesztés során három alapelvet követtek a szerkesztők:

- Vonzó megjelenés és multimedialitás: vizuálisan megnyerőnek lenni mindkét célközönség számára audio- és videoanyagok segítségével.
- Interaktivitás: különösen a diákokra hasson ösztönzően a honlap, ehhez a feladatok különösen fontosak.
- Felhasználóbarát kivitel: a tanárok kész tananyagokat találjanak, amelyeket szabadon felhasználhatnak óráikon.

Tehát a http://languagesindanger.eu/ honlapon angol, holland, német, lengyel és magyar nyelven elérhető szövegek fókuszában elsősorban a veszélyeztetett nyelvek állnak. A közvetített ismereteket hang- és képanyagok, kvízkérdések, feladatok, interaktív térkép, valamint tananya-

gok (óravázlatok) egészítik ki. Érdemes külön is kiemelni azt az enciklopédikus egységet, mely „Tudástár” címen az oktatásban háttér-, illetve kiegészítő anyagként használható ismereteket tartalmaz.

A magyar nyelvű oldal oktatásra szánt részei illeszkedtek a magyar oktatási rendszer akkori adottságaihoz, tartalmi szempontból a Nemzeti Alaptantervhez (NAT), és a kerettantervből indultak ki.

A munkálatok előtt középiskolások körében készítettünk egy előzetes kérdőíves felmérést arról, hogy milyen a tájékozottságuk a nyelvi diverzitás, nyelvi veszélyeztetettség stb. témakörökben. A kérdőívekre adott válaszok kiértékelése megerősített minket abban, hogy a középiskolás tananyagból és a közgondolkodásból hiányoznak ezek az ismeretek.

A tananyagok készítése közben folyamatosan konzultáltunk tapasztalt, különböző típusú iskolákban oktató, budapesti és vidéki középiskolai tanárokkal. Részvételükkel két alkalommal is jó hangulatú workshopot tartottunk. A pedagógus kollégák számos gyakorlati tanáccsal láttak el minket az anyagrészek véglegesítése előtt. Nekik is köszönhető, hogy az elkészült tananyagokban friss tudományos eredmények jelennek meg középiskolások számára könnyen befogadható formában.

2. A honlap felépítése

A weblap – a kötelező információs oldalakon kívül – négy nagy részből áll: „Interaktív térkép”, „Tudástár”, „Tananyagok” és a „Mit tehetsz te?” című egység.

A kezdőlapról elérhető Nyelvek listájában[1] megtaláljuk az oldalon szereplő mind a 201 nyelvet. A betűrendes felsorolásban jellemzően a veszélyeztetettség mértékéről, a nyelv eredet szerinti besorolásáról olvashatunk, helyenként szemléltető videókkal, hangfelvételekkel vagy további információkat tartalmazó oldalakra mutató linkekkel.

[1] http://hu.languagesindanger.eu/nyelvek-listaja/

1. kép. A „Veszélyben a nyelvek” honlap nyitóoldala.

2.1. Tudástár

A „Tudástár” 12 fejezetből áll. Ebben a nyelvészet különböző területei (pl. nyelvleírás, történeti összehasonlító nyelvészet, nyelvtipológia, szociolingvisztika) mellett megjelennek más tudományágaknak – szociálpszichológia, szociológia, jog, kultúrtörténet, kultúrantropológia, klimatológia, gazdaságföldrajz stb. – a nyelv szempontjából fontos aspektusai. A „Tudástár” összefoglalja a magyar nyelven nem elérhető alapismereteket, és számos érdekességet is közöl a következő témakörökben:

1. Az emberi nyelv
2. Hány nyelv van a világon?
3. Hogyan hasonlítjuk össze és írjuk le a különböző nyelveket?
4. A nyelv hangjai
5. Szó – jelentés – szókincs
6. Írás
7. Kultúra és nyelv
8. Többnyelvűség és nyelvi érintkezés
9. Nemzetiségi hovatartozás és nyelv
10. Nyelvi veszélyeztetettség
11. Nyelvpolitika és nyelvtervezés
12. Nyelvi dokumentáció

2.2. Interaktív térkép

A tananyagokban felbukkanó nyelvek közül két tucat egy világtérképen is megjelenik. A pontokra kattintva alapvető ismeretek szerezhetők az adott nyelvről és beszélőiről. Ezt követően érdekes feladatok közül választhatunk, amelyek megoldásával fogalmat alkothatunk az adott nyelvről, valamint találhatunk kapcsolódó fényképeket és videókat is.

2.3. Tananyagok

Itt találhatók az elsősorban tanórákra szánt anyagok: részletesen kidolgozott óravázlatok feladatokkal. Fő céljuk a tartalomközvetítés mellett a különböző kompetenciák fejlesztése, úgymint a kritikus gondolkodás, a problémamegoldás, az együttműködés, a döntéshozatal, az érzelmek kezelése, a kapcsolati kultúra, a társas tolerancia. Fontos volt továbbá a diákok nyelvi horizontjának szélesítése, valamint az érzékenyítés, és ezen keresztül attitűdjük formálása is.

2.4. Mit tehetsz te?

Ennek a résznek a legfőbb feladata a kisebbségi és veszélyeztetett nyelvek iránti érzékenyítés. Ehhez videók, fotók és feladatok nyújtanak betekintést a nyelvészeti munkába, melyeknek a kedvcsinálás is határozott célja.

3. A magyar változatról

A mintául szolgáló anyagokat a konzorciumot vezető lengyel fél készítette el. Ezeket adaptálta a a többi partner, így a magyar változat is részben átdolgozás és sok esetben bővítés eredménye. Nem ragaszkodtunk teljes tartalmi egyezéshez, igyekeztünk a magyar sajátosságokhoz igazítani a „Tudástár” és a „Tananyagok” részeket. Ez azt jelenti, hogy a nagyvilágban található, változatos nyelvi sajátságokat felmutató, és különböző mértékben veszélyeztetett nyelveket beszélő közösségek mellett megjelennek a magyarországi kisebbségek, a határon túli magyar közösségek. A nyelvrokon népekre azért is helyeztünk különösen nagy hangsúlyt, mert e többségükben kicsiny és távoli kultúrák alig ismertek Magyarországon, és a magyar nyelv rokonsága, illetve a nyelvrokon népek iránt érzett, ideológiai okokra visszavezethető ellenszenv nehezíti a róluk szóló ismeretek terjesztését.

A magyar nyelvű „Tudástár”-ban több fejezetet találunk, mint az angol és a lengyel változatban, mert a szövegeket több esetben bővítettük, fejezeteket csoportosítottunk át, és a tematikát is módosítottuk – természetesen az eredeti interaktív és multimédiás jelleg megtartásával.

A „Tananyagok” részben eredetileg négy tantárgy szerepelt: angol mint idegen nyelv, földrajz, társadalomismeret és kulturális ismeretek. A magyar változat számára készült egy további, nyelvtanórákon használható tananyag is. Ez módot adott arra, hogy a magyar nyelv különlegességét, ritka „képességeit”, egyedülállóságát hirdető áltudományos nézetekkel szemben demonstrálni lehessen a világ nyelveiben felfedezhető változatos nyelvi jelenségeket. A feladatok szabadon felhasználhatók a fentieken túl történelem (állampolgári ismeretek), erkölcstan, ének, a hazánkban oktatott kisebbségi nyelvek vagy az osztályfőnöki órákon is. A tananyagok nagy súlyt fektetnek arra, hogy az órai feldolgozás során a diskurzus ne a létező társadalmi problémák és feszültségek megvitatását kezdeményezze, hanem a kisebbségek mai szociális-gazdasági helyzetének történeti, gazdaságtörténeti okait ismerjék meg a diákok. A nyelvi jogokról szóló tananyagnak komplex szerepe van: mivel Magyarországon is számos, törvényben elismert nemzetiség van, viszont a diákság általában nem ismeri ezeket, e tananyagok alkalmasak arra, hogy általában véve is felkeltsék az érdeklődést irántuk. Egyrészt enciklopédikus tudást közvetítenek, másrészt felhívják a figyelmet a többségi nyelvet beszélők felelősségére.

További fontos szempont volt, hogy a nyelvi veszélyeztetettséget olyan témákkal összefüggésben is bemutassuk, amelyekről manapság a közbeszédben is sokat hallani. Ilyen például a globális felmelegedés és az emögött álló iparosítás, bányászati tevékenység.

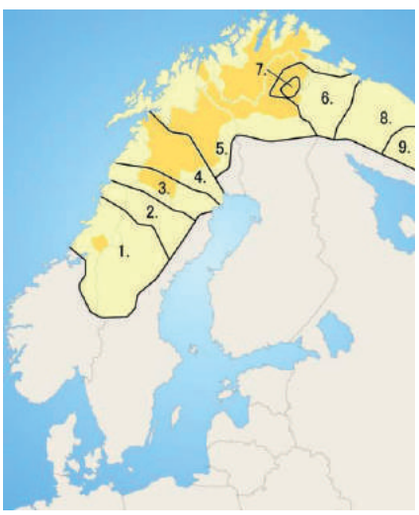
A tematikus sokszínűség mellett a legkülönbözőbb feladattípusokkal próbáltuk segíteni a pedagógusok munkáját: szövegértési és szövegalkotási feladatok, irányított, internetes anyaggyűjtés, csoportos interjúkészítés (például többnyelvű családokkal), a kisebbségi helyzetet megragadó statisztikai jellegű ismeretek feldolgozása stb. Igyekeztünk a feladatokat úgy összeállítani, hogy a tanulók új ismeretek szerzése közben közvetlen, személyes benyomást is szerezzenek olyan témákról, mint a kisebbségi helyzet, kisebbségi oktatás, kétnyelvűség vagy kódváltás – ezzel is növelve társadalmi és szociális érzékenységüket. Így például az egyik feladatban egy szöveg meghallgatása után állításokat kell csoportosítani.

A felvételeken kétnyelvű adatközlők vallanak arról, hogy mit jelent számukra a saját kétnyelvűségük, hogyan boldogulnak az egyik, illetve másik nyelven. Ezáltal a tanulóknak összetettebb képe lesz a többnyelvűségről.

4. Néhány példa

Az alábbiakban azt szemléltetjük, hogy az ismeretterjesztő fejezetekben és a tananyagokban olvasható érdekességek sosem öncélúak, nem a puszta szórakoztatás kedvéért szerepelnek a szövegben. Sok esetben önmagukon túlmutatnak, azaz olyan példák vagy adatsorok, amelyek általános jelenségeket tesznek érthetővé és könnyen megjegyezhetővé.

A világ nyelveit bemutató fejezet körültekintően tárgyalja azt, hogy a nyelvek számának megállapítása többféle problémába ütközik. Ezek egyike, hogy a nyelvek és nyelvjárások elhatárolása nem egyszerű. Ha a politikai aspektusokat ezúttal figyelmen kívül hagyjuk, sok esetben akkor is nehéz eldönteni, hogy nyelvjárási különbségekről van szó, vagy két különböző nyelvről kell beszélnünk. Az átlagember vélhetőleg hallott már a számi (régebbi elnevezése: lapp) nyelvről, és arról is, hogy számos nyelvjárása van. Ezeket a dialektusokat a kölcsönös megértés hiánya miatt a nyelvészek egy része különálló nyelvnek tekinti. Ebben az esetben azonban nem egy, hanem kilenc számi nyelvvel kell számolnunk, vagyis nagy különbségeket eredményezhet a nyelvészeti szempontok érvényesítése.



2. kép. A számi nyelvek Norvégia, Svédország, Finnország és részben Oroszország térképén.[2]

[2] http://commons.wikimedia.org/wiki/File:Sami_languages_large_2.png

A számi példának van egy másik kontextusa is. A térkép jól mutatja, hogy az erős dialektális megoszlást nagyrészt Skandinávia hegyvonulatai okozzák, vagyis azért van nagy különbség a nyelvjárások között, mert a hegyek miatt nehéz volt a kapcsolattartás a csoportok között. Azonban nem csak a hegyek jelenthetnek földrajzi akadályt. Amint földrajz tananyagunkból kiderül, a hanti nyelv helyzete a számihoz hasonlít, de míg ott a hegyvonulatok nehezítik meg a közlekedést, addig a hantiknál a mocsarak.

Más körülmények között hatalmas területen is egységes maradhat a nyelv. A hantik és a nyenyecek egyaránt Oroszország északi vidékein élő népek, és lélekszámuk is hasonló. Míg a hanti nyelv híres nyelvjárási tagoltságáról, addig a nyenyec sokkal egységesebb képet mutat – annak ellenére, hogy beszélői szintén nagy területen élnek. A hantik délebbre, a tajga övezetben, az Ob és mellékfolyói mentén vadászó-halászó életmódot folytattak. A mocsaras, lápos vidéken nem tudtak akkora távokat megtenni. Így életmódjuk ehhez igazodott és csoportjaik egy-egy folyó mentén telepedtek le. A csoportok nem érintkeztek folyamatosan egymással, ezért különültek el a nyelvjárások. Ezzel szemben a nyenyecek északabbra élnek, rénszarvasaikkal a tundra övezethez tartozó hatalmas és gyakorlatilag fátlan területet járják be. Az éghajlat, a terület sajátossága és az ezekhez alkalmazkodó nomád életforma miatt a nyenyecben nem alakultak ki nagy nyelvjárási különbségek.

A nyelvet mint értéket a szövegek nagyon sok oldalról körüljárják. A nyelvnek identitás-, információ-, valamint kultúrahordozó szerepe egyaránt lehet. A nyelv és kultúra kapcsolódását illusztrálják a különböző találós kérdések. Ezek virágzásának nem a városias, írásbeliségen alapuló életforma kedvez, ezért mára már inkább gyerekeknek szóló műfajnak tekintik. A találós kérdések egy jelentős hányada – amelyeknek általános emberi tapasztalat az alapja – komoly hasonlóságot mutat egymással a világ bármely táján, pl. ÁRNYÉK, TOJÁS, FOGAK A SZÁJBAN stb. A következő közmondás bárhol keletkezhetett volna: „Mindkettőnek ötöt fiacskája van, de mint senki más, úgy felpofoznak.” (kéz; spanyol). A találós kérdések másik része erősen kötődik ahhoz a kultúrához, életmódhoz, természeti környezethez, amelyben a nyelv beszélői élnek. Így az alábbi hanti találós kérdés keményebb dió: „Tágas hely lyukas rénbőrrel befedve.” Megfejtéséhez tudnunk kell, hogy az obi-ugorok kúpsátrainak téli borítása rénbőr volt, amelyeken a bögölyszúrások miatt kicsiny lyukak lehettek. Az ezeken keresztül beszűrődő fények a csillagos égre hasonlítottak.

3. kép. A réncsordák télen az eleségért, nyáron a vérszívó rovarok elől menekülve óriási területeket járnak be.[3]

A nyelvi veszélyeztetettségnek külön fejezetet áldoz a „Tudástár". Megtudhatjuk, hogy a beszélők száma ugyan az egyik legfontosabb tényező, de emellett számos egyéb faktor is szerepet játszhat abban, hogy egy beszélőközösség meg tudja-e őrizni nyelvét. A számok relatív voltát szemlélteti az izlandi és a yemba (a niger-kongói nyelvcsalád egy tagja) esete. Ezek a nyelvek körülbelül azonos számú beszélővel rendelkeznek, mégis élőnek és életképesnek tekintjük az egyiket (az izlandit), míg a szintén kb. 300 000 ember által használt afrikai nyelv veszélyeztetettnek számít, mert a mindennapokban az angol és egy helyi pidzsin változat egyre nagyobb mértékben veszi át a helyét. Az uráli vagy finnugor nyelvek közül több is az erősen veszélyeztetett vagy kihalófélben lévő példák között tűnik fel (ilyen a már ötszáz beszélőt sem számláló nganaszan vagy a lív, melynek kihalását többször is bejelentették már).[4] Üdítő kivétel az észt, mely a világ nyelvei között nem számít nagynak a maga közel egymillió beszélőjével, mégis az egyik legéletképesebb uráli nyelv.

[3] http://commons.wikimedia.org/wiki/File:Reindeer-on-the-rocks.jpg
[4] https://www.nyest.hu/renhirek/elhunyt-grizelda-kristin

A nyelvtervezésnek fontos szerepe lehet egy nyelv megmaradásában vagy eltűnésében. Jól átgondolt oktatáspolitikával kisebbségben élő nyelvek beszélői is hosszú távon biztosan őrizhetik nyelvüket, de az iskola lehet a pusztítás terepe is. Feltehetően többek által ismert tény, hogy az amerikai indián nyelvek kihalásában a misszionáriusok által vezetett „átnevelő" iskoláknak is igen nagy volt a szerepe. Az viszont kevésbé köztudott, hogy a múlt század harmincas éveitől kezdve a Szovjetunióban is hasonló folyamat zajlott le. Az elszórt kis településekről központi internátusokba gyűjtötték a gyerekeket, akik nemcsak családjuktól elszakítva töltötték a tanév nagy részét, hanem a közös nyelv jellemzően az orosz volt, így igen gyorsan elszoktak anyanyelvük használatától.

Ma már más problémákat vet fel az anyanyelv oktatása. A kisebb uráli nyelvek esetében az iskolában ugyan heti néhány órában lehet foglalkozni a nyelvvel, azonban sok szülő szerint ez felesleges időtöltés, mert nem segíti később az egyéni boldogulást.

4. kép. Nganaszan diákok Dugyinkában tanárnőjükkel.
Fotó: Várnai Zsuzsa.

Az anyanyelvünkhöz tartozó kultúrát nemcsak a családban és az iskolában ismerhetjük meg, fontos szerepe van a közösség által őrzött hagyományoknak és a továbbadás lehetőségének. Például a kelta nyelvek közül a walesi esetében az Eisteddfod (bárdköltészeti) fesztiválnak[5] is jelentős

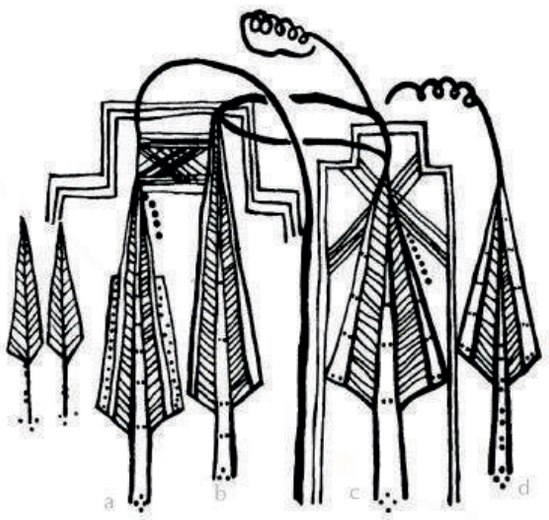
[5] https://eisteddfod.wales/gallery/2018-cardiff

szerepe van a beszélőközösség nyelvmegőrzésében. Szemben az Írországban, Skóciában és Bretagne-ban is tapasztalható helyzettel, ahol a kulturális azonosságtudat szintén megvan, ez azonban mégsem elegendő a kelta nyelvű beszélők nagyarányú nyelvcseréjének megakadályozásához.

Mindannyian ismerjük azt a városi legendát, amely szerint az eszkimó nyelvben száz körüli szó használatos a hóra. Erről ugyan kiderült azóta, hogy erős túlzás, mégis felhívja a figyelmet arra a jelenségre, hogy a nyelvek valóban jelentős különbségeket mutathatnak a világ jelenségeinek a megnevezésében. A „Tudástár"-ban számos hiteles példa olvasható erre. Az egyik dél-afrikai nyelvben a magyar *visz* igének legalább öt megfelelője van attól függően, hogy melyik testrészen van a súly (fej, kar, váll), illetve hogy a cselekvő mit visz (terhet, gyereket). A közép-afrikai baka nyelv nyolc szót használ az elefántra. Mivel az elefántvadászat az egyik legfontosabb tevékenységük, az egyes példányok nemét, korát, méretét, erejét is kell tudniuk; valamint az elnevezés még arra is utalhat, hogy magányos vagy csordában élő állatról van-e szó. A fajon belüli differenciálás általában az életmóddal van összefüggésben: a réntartó népek a rénszarvasokra, a lótartók a lovakra alkalmaznak nagy számú megnevezést. Emellett a biológiai diverzitás is leképeződhet a nyelvben: az óceánok mellett lakó népek a körülöttük élő halfajokat igen jól ismerik. A helyi nyelvek nemegyszer a biológusokat is segítik a fajok, alfajok azonosításakor, így az őshonos nyelvek szavai megjelenhetnek a latin nyelvű elnevezések részeként is.

Ha írásról van szó, elsősorban betűírás, szótagírás vagy képírás jut eszünkbe. Arra ritkábban gondolunk, hogy az írásrendszerek létrejötte előtt vagy azok hiányában is érezhették szükségét az emberek, hogy gondolataikat megörökítsék vagy ne közvetlenül közöljék. Nagyon látványosak például a jukagír szerelmeslevelek, amelyek valójában nem tekinthetők írásnak, amennyiben a jelek nem kapcsolhatók semmiféle nyelvi egységhez vagy struktúrához, hanem közvetlenül utalnak fogalmakra. Az 5. képen látható szerelmeslevélben az emberek fenyőkhöz hasonló figurák, körülöttük keretszerűen jelenik meg a ház, a házban egymást keresztező vonalak konfliktusokra utalnak; a szerelmi sokszögben a lány szereplőket pontsorral jelölt hajfonat különbözteti meg, magát a szerelmi érzést pedig a fák tetejéről induló kacskaringós vonal mutatja. A kis fák gyerekeket jelölnek, és mivel nem a családot is jelképező házon belül vannak, tudjuk, hogy még nem születtek meg. Az ábra mondandója a következő: a jukagír lány szerelme egy orosz nővel él együtt, akivel

kapcsolata nem békés. A szerelmes lány érzései igen erősek. Magányában a lányt egy másik fiatalember ostromolja szerelmével, ezért az orosz nővel élő fiúnak hamar kell választani, mielőtt gyerekeik születnének.



5. kép. Egy jukagír szerelmeslevél (Sampson, 1985: 28–29).

Különböző nyelvekkel találkozunk nap mint nap a körülöttünk lévő feliratokon is. Mára már önálló tudományterületté fejlődött a „nyelvi tájkép” kutatása. A honlap a magyar diákok számára is megvilágítja a többnyelvűség megjelenítésének fontosságát a feliratokon. Az alábbiakban azokból a fotókból adunk válogatást, melyek a tárgyalt problémákat illusztrálják.

6. kép. Montázs a honlap „nyelvi tájkép” fotóiból.

Végezetül a „Mit tehetsz te?” részből a lengyel Tymoteusz Król példáját említjük, aki 11 éves korában kezdte el dokumentálni nagymamája nyelvét. A kihalófélben lévő wilamowiceiről van szó, melyet az UNESCO 2007-ben Tymoteusz javaslatára sorolt a súlyosan veszélyeztetett nyelvek közé. Az információhoz a következő feladat társul:

> „Nézd meg a videót, melyben Tymoteusz (immár fiatalemberként) dokumentációs tevékenységéről beszél. Az interjú után hallhatod Rozalia Hanuszt (1926–2009), amint Wilamowice településről beszél (wilamowicei nyelven).”[6]

5. Befejezés

A már befejeződött projekt létjogosultságát semmi nem igazolja jobban, mint hogy időről időre találkozunk olyan hivatkozásokkal, melyek azt bizonyítják, hogy az anyagot eredeti szándékunknak megfelelően használják a középiskolások, egyetemisták és tanáraik. Az általunk készített anyagok részletei szerepelnek a Magyar Nyelvészeti Diákolimpia felkészítő anyagai között éppúgy, mint egyetemi kiselőadások, középiskolai projektfeladatok irodalomjegyzékeiben. Visszagondolva a közös munkálatok jó hangulatára is, szívesen állunk hasonló kihívások elé a jövőben.

7. kép. „Ordítozás” az intézet előadótermében.

[6] http://hu.languagesindanger.eu/what-can-be-done/dokumentalj-egy-nyelvet-vagy-nyelvjarast/#probald-ki:

Bibliográfia

Duray Zs., Várnai Zs.: Az INNET-projektről: Nyelvi veszélyeztetettség oktatása középiskolában: Idegen nyelv – anyanyelv. Édes Anyanyelvünk 36/3, 15 (2014)

Duray Zs., Oszkó B., Sipos M., Szeverényi S., Várnai Zs.: INNET: Nyelvi veszélyeztetettség, nyelvi kisebbség, nyelvi diverzitás fogalmak bevezetése a magyar közoktatásba. In: Szöllősy É., Prax, L.; Hoss, A. (szerk.) Találkozások az anyanyelvi nevelésben. pp. 54–69. Pécsi Tudományegyetem Nyelvtudományi Doktori Iskola, Pécs, Magyarország (2013)

Jung, D., Klessa, K., Duray, Zs., Oszkó, B., Sipos, M., Szeverényi S., Várnai Zs., Trilsbeek, P., Váradi T.: Languagesindanger.eu – Including Multimedia Language Resources to disseminate Knowledge and Create Educational Material on less-Resourced Languages. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 530–535. European Language Resources Association (ELRA), Reykjavik, Iceland (2014)

Sampson, G.: Writing Systems. A linguistic introduction. Hutchinson, London (1985)

Sipos M.: Érdekességek a nyelvről és a nyelvekről az INNET-projekt honlapján: internetes kiegészítő anyagok a nyelvtan tanításához. Anyanyelv-pedagógia 7/4 15 p. (2014) http://www.anyanyelv-pedagogia.hu/cikkek.php?id=540

Szeverényi S., Sipos M.: Az Innet-projekt a nyelvi sokféleségről. Édes Anyanyelvünk 36/1, 11 (2014)

Hangok, hangulatok, gesztusok: magyar nyelvű dialógusok multimodális vizsgálata

Hunyadi László[1], Szekrényes István[2]

[1] Debreceni Egyetem, Általános és Alkalmazott Nyelvészeti Tanszék
hunyadi@unideb.hu
[2] Debreceni Egyetem, Filozófia Intézet
szekrenyes.istvan@arts.unideb.hu

1. Bevezetés

Egy, a Nyelvtudományi Intézettel, közelebbről annak számítógépes nyelvészeti osztályával és még közelebbről Váradi Tamással való évtizedes együttműködés eredménye az a HuComTech korpusz építésében és vizsgálatában megtestesülő, majd további irányokba vezető, 2009-ben indult kutatássorozat, amelynek elsődleges célja a multimodális kommunikáció vizsgálata elméleti alapjainak (Hunyadi, 2011; Németh, 2011) lefektetése volt. E munka sokszínűségét és az együttműködések sokféle szintjét jellemzi, hogy a Debreceni Egyetemen bölcsész-, informatikai és mérnöki karain és az MTA Nyelvtudományi Intézeten kívül részt vettek benne még a BME, az MTA TTK Pszichológiai Kutatóintézet, a Szegedi Tudományegyetem és a National Instruments Hungary szakemberei, sőt, egyes, mindezzel érintkező további kutatásokban-fejlesztésekben a Debreceni Egyetem orvosai és a Miskolci Egyetem mérnökei is.

A kutatás-fejlesztés célját természetesen jelentősen meghatározta, hogy a hazai (és nemzetközi) innovációs igényeknek az egyre hangsúlyozottabb előtérbe kerülésével nyilvánvalóvá vált, hogy bizonyos konkrét fejlesztésekhez egymástól látszólag távol eső szakterületek, tudományterületek, sőt szakmák együttes munkájára van szükség, határaikat egymáshoz kell közelíteni és átjárhatóvá tenni. Ehhez jó kiindulásnak bizonyultak a számítógépes nyelvészek, akik már addig is folyamatos dialógust folytattak szoftverfejlesztőkkel és mérnökökkel, de ugyancsak a mérnökök, akik beszédfeldolgozó algoritmusokon dolgoztak, vagy megsejtették, hogy az általuk épített robotoktól emberszerűbb viselkedést várnak a felhasználók. A pszichológusok is lelkesen csatlakoztak, hiszen az érzelmek, szándékok kutatása, beleértve azok felismerését és

DOI: https://doi.org/10.18135/VT70.10

szerepük tanulmányozását a kommunikáció sikerében vagy sikertelenségében, fontos elméleti és gyakorlati jelentőséggel bír számukra is. A nyelvész-pragmatikusokat mindenekelőtt a konverzáció folyamatának nyelvi vetülete érdekelte, kiegészítve mindezt a gesztusok és azok funkcióinak a tanulmányozásával, így multimodálissá szélesítve az addigi hagyományok verbálisközpontúságát. A fonetikusoknak lehetőségük nyílt arra, hogy informatikusokkal karöltve új algoritmusokat dolgozzanak ki a prozódia által közvetített tartalom felismerésére, egyebek között a gépi tanulás módszereivel feltárva a multimodális jelek közötti jellemző funkcionális összefüggéseket. A szintaxis kutatói is új lehetőségekhez juthattak azáltal, hogy a korpusz automatikus mondattani elemzése által első ízben kaptak lehetőséget a beszélt nyelv szintaxisának az eddigieknél átfogóbb igényű és terjedelmű megragadására. A bizonyos részleteiben egyre szélesedő kutatási palettán végül megjelentek fül-orr-gége szakorvosok is, akik a siketekkel való kommunikáció lehetőségeinek a bővítését tűzték ki célul a beszédakusztika elérhetővé tételével nem hallók számára, valamint annak az artikulációval való újszerű összekapcsolásával (Hunyadi és mtsai., 2015).

2. Az adatgyűjtés módszerei

Több hónap tervezés és előkészítő tevékenység után a kutatássorozat empirikus forrását jelentő HuComTech korpusz (Hunyadi és mtsai., 2012, 2016c) hang- és videóanyagát 2010 tavaszán, a Debreceni Egyetem Angol–Amerikai Intézetének stúdiójában, 111 (54 nő és 67 férfi, átlagéletkor: 22 év), főként egyetemista korú beszélő közreműködésével készítettük el. A mintegy 50 órát kitevő felvételanyagon 222 interjúbeszélgetést rögzítettünk, amelyek részben az adatközlőkkel készített szimulált állásinterjúkból, illetve az ezeket követő informális beszélgetésekből tevődtek össze. Az utóbbi esetében az interjúvezető egy előre kidolgozott kérdéssor segítségével, eltérő érzelmi töltetű reakciók kiprovokálásával (pl. „Kérlek, mesélj egy negatív élményről, amit mostanában átéltél!”) adott keretet a dialógusnak. A beszélgetések túlnyomó többségét ugyanazon személy vezette. Ennek előnye, hogy – egy esetleges további kutatás céljából – adott a lehetőség egyebek között a beszélgető partnerhez való sokféle alkalmazkodás vizsgálatához is.

A résztvevők ülő helyzetben történő beszélgetésének hanganyagát 2 darab Shure 16A típusú mérőmikrofon segítségével, 44 100 Hz-es mintavételezési frekvencia és 16 bites kvantálás mellett 2 csatornán rögzítettük, az annotáláshoz és az akusztikai elemzéshez később a felvételek egy

csatornára mixelt verzióját használtuk fel. Az interjúk képanyagát 3 pozícióból (2 kamerát irányítottunk az adatközlőre, egyet pedig az interjúvezetőre) nagy felbontásban vettük fel, 3 darab Sony HDRXR520VE típusú, statikus állványokra helyezett kamera használatával. A felvételeken a beszélők térdtől felfelé láthatóak.

A hangfelvételek elemzéséhez a *Praat* program (Boersma és Weenink, 2020) annotációs funkcióját használtuk, amely egy szöveges formátumú, más beszédtechnológiai platformok által is könnyen importálható és feldolgozható kimenetet produkál. A videófelvételek annotálásához a DE ITK Képfeldolgozó Csoportja *QANNOT* néven fejlesztett egy saját alkalmazást (Pápay és mtsai., 2011), amely lehetővé tette a felvételek képkockáról képkockára történő, gördülékeny címkézését. A program az elemzéshez használt kategóriákat és a választható értékek hierarchikus szerkezetét egy külső XML-állományból dinamikusan olvasta be, amelynek elkészítése, illetve más annotációs feladatokra való átdolgozása, majd később a címkéket rendszerező relációs adatbázis struktúrájának kialakítása megkívánta a bölcsész kollégákkal való folyamatos konzultációt és egy közösen értelmezhető terminológia kialakítását.

A korpusz felvételeinek alapszintű annotálása mintegy két évet és egy tucatnyi annotátor együttes munkáját vette igénybe, ami magában foglalta a beszéd és a speciális beszédesemények (hezitáció, nevetés, levegővétel stb.) standard jelölésékkel történő leiratozását, az érzelmek, a fordulóváltások és a nonverbális gesztusok címkézését (Pápay és mtsai., 2011). Mindez később (további 6 év munka után) kiegészült a teljes szöveg fonetikai, morfológiai és szintaktikai leírásával, a dialógusok pragmatikai elemzésével és a prozódia automatikus annotálásával is.

Az automatikus morfológiai és szintaktai elemzéshez a Szegedi Tudományegyetemen fejlesztett *magyarlanc* (Zsibrita és mtsai., 2013) alkalmazás kimeneteit használtuk fel. A korpusz teljes anyagát lefedő, speciális kódolási sémát alkalmazó manuális elemzés pedig Kiss Hermina munkájának köszönhető (Kiss, 2014). A multimodalitás mint alapvető szempont érvényesítése érdekében a CLARIN-D projekt WebMAUS (Kisler és mtsai., 2017) szolgáltatásával elkészítettük a korábban csak a megnyilatkozások és a tagmondatok szintjén szegmentált szöveg szószintű időillesztését, amivel lehetővé válik az egyes szavak, kifejezések vagy mondatok más, akár nem nyelvi attribútumokhoz (pl. gesztusokhoz, pragmatikai funkciókhoz), valamint a prozódiához való illesztése is.

A nem verbális kommunikációs szintek közül annotáltuk az arc, a tekintet, a felsőtest, a fej és a kéz mozgásait, ezekhez fizikai jellemzőket

(pl. mozgás vagy változás iránya) illesztve, de ugyancsak hozzáadva az érzelmi és pragmatikai attribútumokat is. A megfigyelő által értelmezett érzelmeket annotáltuk multimodálisan a hang és a videó együttes érzékelésével és unimodálisan is, egyedül a hang alapján. A sokrétű pragmatikai annotálásból, amely magában foglalt minden lényeges és hagyományos, szövegalapú jellemzőt (beszédváltás, különböző beszédaktusok, új és régi információ) újdonságként kiemeljük a beszélés elkezdésének (videóban és/vagy hangban érzékelhető) szándékát, ami nem feltétlenül esik egybe a beszélés valóságos kezdetével.

A beszéddallam automatikus elemzését egy saját fejlesztésű, a *Praat* program szkriptnyelvén implementált algoritmus (Szekrényes, 2014, 2015) segítségével végeztük el. A fejlesztés során arra törekedtünk, hogy az intonáció perceptuálisan releváns változásait az alapfrekvencia-görbe nagyobb dallamtrendekre történő stilizálásával, szegmentálásával és a beszélő egyéni sajátosságaihoz adaptált kategorikus címkézésével ragadjuk meg. A később XML-formátummal és vizuális megjelenítésre alkalmas XSL-stíluslapokkal is kiegészített, eredetileg *Praat TextGrid* formátumú kimenet a mért értékek mellett számot ad a dallamszegmentumok különböző karakteréről (pl. „emelkedő", „eső", „szinttartó"), illetve a beszélő 5 tartományra felosztott hangterjedelmében elfoglalt relatív pozíciójáról. A módszert később kiterjesztettük az intenzitás és a beszédtempó hasonló céllal történő vizsgálatára is. A beszéddallam elemzését végző eljárás később az *e-magyar* projekt (Váradi és mtsai., 2017) keretében, *emPros*[1] néven vált részévé egy nyílt forráskódú megoldásokat adoptáló beszédelemző lánc moduljainak (Kornai és Szekrényes, 2017). Itt az *e-magyar* projektet vezető Váradi Tamás és a beszédfeldolgozó alprojektet irányító Kornai András érdemeként kell kiemelnünk, hogy a korpuszban tárolt adatok mellett egy, addig csak belső használatra szánt automatikus eljárás is publikusan elérhető vált az érdeklődő szakmai közönség számára. Az algoritmus flexibilitásának javításában előzetesen nagy segítséget jelentettek a *SegCor* projekt[2] munkatársai is, akik lehetővé tették a FOLK korpusz (Schmidt, 2016) hangfelvételein való tesztelést.

[1] http://e-magyar.hu/hu/speechmodules/empros
[2] https://segcor.cnrs.fr/

3. A korpusz közzététele

A több millió címke lejegyzése önmagában korlátozott jelentőséggel bír, ha – mivel a kommunikáció alapvető tulajdonsága, hogy időben zajlik – a címkékben és kapcsolataikban hordozott információ nem kereshető vissza és nem elemezhető más címkék jelenlétének/hiányának időbeli függvényében. Az első adatbázis, amelyet az adatok elemzéséhez építettünk, SQL-alapú volt, amely így lehetővé tette standard SQL-lekérdezések alkalmazását. Az adatelemzésnek ez a módszere azonban még a lekérdezéshez készített grafikus interfésszel is megkövetelte a felhasználótól az adatbázis struktúrájának, a mögöttes technikai megoldásoknak a pontos ismeretét, ezért csak a szűkebb kutatócsoporton belül tudtuk hasznosítani. A tágabb kutatóközönség kiszolgálásához más módszerek alkalmazására, a metaadatok standard formában történő rögzítésére volt szükség.

Váradi Tamás a CESAR (Váradi, 2012) és a CLARIN projekt hazai koordinátoraként, később a HunCLARIN megalapítójaként szerzett szakmai tapasztalatai és javaslatai a kutatómunka ezen fázisában hatalmas segítséget jelentettek. A Nyelvtudományi Intézettel a CESAR projekt keretében folyó újabb együttműködés keretében a korpusz metaadatait és XML-formátumba konvertált állományait először a META-SHARE online felületén tettük közzé. Később a *The Language Archive* projekt[3] (a továbbiakban: TLA) által preferált IMDI metaadatsémára (Broeder és Wittenburg, 2006) áttérve az ARBIL program (Withers, 2012) és a LAMUS (Broeder és mtsai., 2006) rendszer használatával minden adatot elérhetővé tettünk a TLA nijmegeni gyűjteményében, ahol bárki hozzáférhet a korpuszhoz a TROVA kereső[4] és az ANNEX (Berck és Russel, 2006) lekérdezőfelületen keresztül (a médiaanyagokhoz előzetes engedély alapján). Fontos továbblépés volt, amikor a korpuszban használt annotációs sémákról készített részletes útmutató[5] elkészítése után a TLA HuComTech teljes anyagát tükröztük a Nyelvtudományi Intézet szerverén,[6] a hazai adatbázisok gyűjteményében is, ezzel hozzájárulva a magyarországi nyelvészeti kutatások-fejlesztések szélesköru bemutatásához. Ezen adatokat már eddig is számos munka, köztük eddig két megvédett PhD-értekezés (Abuczki, 2014; Szekrényes, 2020) használta fel, valamint beszédtechnológiai fejlesztéshez is alkalmazták.

[3] https://archive.mpi.nl/tla/
[4] http://tla.mpi.nl/tools/tla
[5] https://tla.nytud.hu/info/hucomtech/guide.html
[6] https://tla.nytud.hu

A hozzáférés továbbra is biztosított szerteágazó kutatások jövőbeli specifikus céljaira. Az adatok elemzésére az ANNEX kereten kívül alkalmas a közismert, szabad hozzáférésű ELAN (Wittenburg és mtsai., 2006) szoftver is, amely az adatfájlokat saját gépre letöltve ugyancsak kényelmes elemzőeszköznek bizonyul.

4. Az adatok elemzése

A korpusz adatait a deskriptív, a különböző modalitások alá tartozó címkék gyakoriságát és együttállásait vizsgáló statisztikák mellett a Nyelvtudományi Intézet és az MTA-SZTE Mesterséges Intelligencia Kutatócsoport bevonásával gépi tanulással végezett kísérletekhez is felhasználtuk. Ezek egy része az interjúkban jelölt témaváltások automatikus, szövegfüggetlen detektálására irányult (Kovács és Váradi, 2017; Kovács és Szekrényes 2019), amit többféle, az annotálás során használt elemzési szintek (a videón megfigyelt nonverbális gesztusok, a megnyilatkozások prozódiai és szintaktikai szerkezete) címkéit összefogó jellemzőcsoport alapján is kipróbáltunk. Egy másik kísérlet az interjúk formális és informális felvételekre történő osztályozását célozta, amelyhez kizárólag a prozódia és a beszélőváltások ritmusát reprezentáló jellemzőkre hagyatkoztunk (Szekrényes és Kovács, 2017). Ezeknek a kutatásoknak és fejlesztéseknek a célja elsősorban egyik esetben sem az adott feladatra maximális hatékonyságot garantáló eljárás kivitelezése volt, hanem az egyes modalitások együttműködésére, informativitására vonatkozó hipotéziseinknek az ellenőrzése.

Az adatok elemzésében jelentős előrelépésnek számít, hogy csatlakoztunk a MASI nemzetközi hálózathoz (Multimodal Analysis of Social Interactions) és használóivá váltunk a *Theme* szoftvernek (Magnusson, 2000). E kifejezetten a viselkedés időbeli multimodális mintázatainak a feltárására létrehozott szoftver válasz más (így idősoros) elemzési módszerek azon korlátozottságára, hogy azokkal szemben képes azonosítani olyan viselkedési mintázatokat is, amelyeknek az egyes elemei csupán opcionálisak, és időbeli jellemzőik (kezdet, vég, időtartam) sem állandóak. A *Theme* szoftver mint kutatási keretrendszer segítségével így számos olyan viselkedési mintázatot sikerült feltárnunk, amelyek – a mintázatok egyes összetevő elemeinek opcionalitása és a figyelembe vett események közötti idő variabilitása miatt – jobbára észrevétlenek maradnának. Így a korpusz egy részkorpuszán leírtuk az élőbeszéd töredezettségének szintaktikai jellemzőit (Hunyadi és mtsai., 2016a), a prozódia és a beszélt szintaxis összefüggéseit (Hunyadi és mtsai., 2016b), az *egyetértés/egyet*

nem értés (Hunyadi, 2019), valamint az *öröm* mint kommunikációs esemény multimodális mintázatait (Hunyadi, 2020). A *Theme* alapján kapott mintázatoknak önállóan, valamint az ELAN annotációs és feldolgozó rendszerben való további vizsgálatára egy SQL-alapú, de könnyen használható webes felületet[7] is létrehoztunk és nyílt felhasználásúvá tettünk (Szekrényes, 2019). A korpusz, köszönhetően komplexitásának és méretének, valamint elérhetőségének, remélhetően még további sokrétű és multidiszciplináris vizsgálatok gazdag lehetőségét fogja nyújtani.

5. Összegzés

A Nyelvtudományi Intézet másutt aligha tapasztalt értékes módon járul hozzá e kutatások kiszélesítéséhez, a HuComTech korpusz adatainak kivételes léptékű feldolgozásához és a módszer szélesebb körökben való elterjesztéséhez: amellett, hogy kezdeményez és helyt ad meghívásoknak, konzultációknak, szakemberek cseréjének, az Intézet a gazdája annak a virtuális számítógéprendszernek is, amely komoly kapacitásával a felhőben végzi adataink feldolgozását.

Bibliográfia

Boersma, Paul, Weenink, David: Praat: doing phonetics by computer [Computer program]. Version 6.1.36, retrieved 6 December 2020 from http://www.praat.org/

Broeder, D., Claus, A., Offenga, F., Skiba, R., Trilsbeek, P., Wittenburg, P.: LAMUS: The language archive management and upload system. In: Proceedings of LREC 2006. pp. 2291–2294 (2006)

Broeder, D., Wittenburg, P.: The IMDI metadata framework, its current application and future direction. *International Journal of Metadata, Semantics and Ontologies* ½. pp. 119–132 (2006)

Hunyadi, L.: Multimodal human-computer interaction technologies. theoretical modeling and application in speech processing. *Arguemntum* 7. pp. 240–260 (2011)

Hunyadi, L., Földesi, A., Szekrényes, I., Staudt, A., Kiss, H., Abuczki, A., Bódog, A.: Az ember–gép kommunikáció elméleti-technológiai modellje és nyelvtechnológiai vonatkozásai. In: Általános nyelvészeti tanulmányok XXIV: Nyelvtechnológiai kutatások. pp. 265–309. Akadémiai Kiadó, Budapest (2012)

Hunyadi, L., Kiss, H., Szekrényes, I.: Incompleteness and fragmentation: Possible formal cues to cognitive processes behind spoken utterances. In: Jeffrey W. Tweedale, Rui, Neves-Silva, Lakhmi C. Jain, Gloria, Phillips-Wren, Junzo Watada, Robert J. Howlett (szerk.) Intelligent Decision Technology Support in Practice. pp. 231–257. Springer International Publishing, Cham (2016a)

[7] https://altnyelv.unideb.hu/ThemeToMySQL/login.php

Hunyadi, L., Kiss, H., Szekrényes, I.: Prosody enhances cognitive infocommunication: Materials from the hucomtech corpus. In Esposito, A., Jain, C. L. (eds.) Toward robotic socially believable behaving systems – volume I: Modeling emotions. pp. 183–204. Springer International Publishing, Cham (2016b)
Hunyadi, L., Váradi, T., Szekrényes, I.: Language technology tools and resources for the analysis of multimodal communication, In: Erhard Hinrichs, Marie Hinrichs, Thorsten Trippel (eds.) Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH 2016). 117–124. University of Tübingen, Tübingen (2016c)
Hunyadi, L., Szekrényes, I., Sziklai, I.: Vizuális percepció és nyelvi feldolgozás. Beszédkutatás 23, 186–208 (2015)
Hunyadi, L.: Agreeing/Disagreeing in a Dialogue: Multimodal Patterns of Its Expression. Frontiers in Psychology 10, 1–9 (2019)
Hunyadi, L.: Happy hour: the multimodal analysis of 'being happy' in a conversation (2020, kézirat)
Kisler, T., Reichel U. D., Schiel F.: Multilingual processing of speech via web services, Computer Speech & Language 45, pp. 326–347 (2017)
Kiss, H.: A HuComTech audio adatbázis szintaktikai szintjének multimodális vizsgálata. In: Tanács, A., Varga, V., Vincze, V. (szerk.) X. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2014) pp. 27–38. Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged. (2014)
Kornai, A., Szekrényes, I.: e-Magyar beszédarchívum. In: Tanács, A. Vincze, V. (szerk.) XIII. magyar számítógépes nyelvészeti konferencia (MSZNY 2017). pp. 103–109. JATEPress, Szeged (2017)
Kovács, G., Váradi, T.: A különböző modalitások hozzájárulásának vizsgálata a témairányítás eseteinek osztályozásához a hucomtech korpuszon. In: Tanács, A., Vincze, V. (szerk.) XIII. magyar számítógépes nyelvészeti konferencia (MSZNY 2017) pp. 103–109. JATEPress, Szeged (2017)
Kovács, Gy.: Classification of Formal and Informal Dialogues Based on Emotion Recognition Features. In: Sojka, P.; Horák, A.; Kopeček, I., Pala, K. (eds.) Text, Speech, and Dialogue: 21st International Conference, TSD 2018, Brno, Czech Republic, September 11–14, 2018, Proceedings. pp. 518–526. Springer Nature, Cham (2018)
Kovács, G., Szekrényes, I.: Applying neural network techniques for topic change detection in the hucomtech corpus. In: Hunyadi, L., Szekrényes, I. (eds.) The temporal structure of multimodal communication: Theory, methods and applications. pp. 147–162. Springer International Publishing, Cham (2019)
Magnusson, M. S.: Discovering hidden time patterns in behavior: T-patterns and their detection. Behavior Research Methods, Instruments, & Computers 32/1, 93–110 (2000)
Németh, T. E. (szerk): Ember-gép kapcsolat. A multimodális ember-gép kommunikáció modellezésének alapjai. Budapest: Tinta Könykiadó (2011)
Pápay, K., Szeghalmy, S., Szekrényes, I.: Hucomtech Multimodal Corpus annotation. Argumentum 7, 330–347 (2011)

Schmidt, T.: Good practices in the compilation of folk, the research and teaching corpus of spoken german. In: Kirk, J. M., Andersen, G. (eds.) Compilation, transcription, markup and annotation of spoken corpora, special issue of the international journal of corpus linguistics [IJCL 21:3] pp. 396–418 (2016)
Szekrényes, I.: Annotation and interpretation of prosodic data in the hucomtech corpus for multimodal user interfaces. Journal on Multimodal User Interfaces 8/2, 143–150 (2014)
Szekrényes, I.: Prosotool, a method for automatic annotation of fundamental frequency. In: 6th IEEE International conference on cognitive Infocommunications (CogInfoCom). pp. 291–296. IEEE, New York (2015)
Szekrényes, I., Kovács, G.: Classification of formal and informal dialogues based on turn-taking and intonation using deep neural networks. In: Karpov, A., Potapova, R., Mporas, I. (eds.), Speech and computer. pp. 233–243. Springer International Publishing, Cham (2017)
Szekrényes, I.: Post-processing T-patterns Using External Tools From a Mixed Method Perspective. Frontiers in Psychology 10, 1–12 (2019)
Szekrényes, I.: Prozódiai jellemzők gépi feldolgozása és hasznosítása élőnyelvi korpuszok elemzésében. PhD-értekezés. Debreceni Egyetem, Nyelvtudományok Doktori Iskola (2020)
Váradi, T.: Central and South-East European Resources in META-SHARE. In: Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012). pp. 431–438 (2012)
Váradi, T., Simon, E., Sass, B., Gerőcs, M., Mittelholcz, I., Novák, A., Indig, B., Prószéky, G., Vincze, V.: Az e-magyar digitális nyelvfeldolgozó rendszer. In: Tanács, A., Vincze, V. (szerk.) XIII. magyar számítógépes nyelvészeti konferencia (MSZNY 2017). pp. 103–109. JATEPress, Szeged (2017)
Withers, P.: Metadata management with Arbil. In: V. Arranz, D. Broeder, B. Gaiffe, M. Gavrilidou, M. Monachini (eds.) Proceedings of the workshop describing LRs with metadata: Towards flexibility and interoperability in the documentation of LR at LREC 2012. pp. 72–75. ELRA (2012)
Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H.: ELAN: a professional framework for multimodality research. In: Calzolari, N. et al. (eds.) Proceedings of LREC 2006. pp. 213–269. ELRA (2006)

Alkalmazott nyelvészet, korpuszok és adatbázisok. A nyelvtechnológia és a korpusznyelvészet a terminológiaoktatásban

Fóris Ágota[1]

[1] KRE BTK Magyar Nyelvtudományi Tanszék és TERMIK.
foris.agota@kre.hu

1. Bevezetés: alkalmazott nyelvészet és tudományszervezés

Váradi Tamással való megismerkedésemet Klaudy Kingának köszönhettem. A Miskolci Egyetemen 1992-ben jött létre a Bölcsészettudományi Intézet (később Kar), ahol Klaudy Kinga az Alkalmazott Nyelvészeti Tanszéket vezette 1992 és 2002 között, továbbá a Modern Filológiai Intézet igazgatója volt. Váradi Tamás követte őt a Modern Filológiai Intézet élén, valamint ő volt az Angol Nyelvészeti Tanszék vezetője 2002 és 2006 között. A XV. Magyar Alkalmazott Nyelvészeti Kongresszust 2005 áprilisában a Modern Filológiai Intézet szervezte Miskolcon, a MANYE-val közösen („A világ nyelvei – a nyelvek világa. Soknyelvűség a gazdaságban, a tudományban és az oktatásban"). A kongresszus előkészületei már 2004 őszén megkezdődtek, amikor a szervezőbizottság Miskolcon tartotta első ülését; Klaudy Kinga, Váradi Tamás, Dobos Csilla a miskolciak részéről, jómagam pedig a MANYE képviseletében vettem részt rajta. A kongresszus mindig nagyszámú résztvevőt vonzott, 300–500 fő között mozgott a jelentkezők száma. Ez a tudományos és az egyéb programok szervezésben és a pénzügyi tervezésben és gazdálkodásban is sok felelősséget és feladatot rótt a szervezőkre. Váradi Tamás a szervezőbizottság elnökeként olyannyira odafigyelt a szervezési feladatokra, hogy végül a miskolci az egyik legsikeresebb kongresszus volt, nyereséges pénzügyekkel. Ezt követően 2005 és 2008 között a MANYE választmányának tagja volt.

Kutatási területe a számítógépes nyelvészet (korpuszfejlesztés, lokális grammatikák, lexikai adatbázisok, gépi fordítás), a szociolingvisztika és az idegennyelv-elsajátítás, továbbá több jelentős magyar szótár munkálataiban vett részt. Nagyszámú publikáció, köztük számos élvonalbeli nemzetközi publikáció szerzője. Az MTMT2 alapján idéző közleményeinek száma több mint 2000.

DOI: https://doi.org/10.18135/VT70.11

Az alkalmazott nyelvészet, ezen belül a nyelvtechnológia, a korpusznyelvészet és a lexikográfia nemzetközileg elismert kutatója, szótárak, korpuszok, adatbázisok, eszközök fejlesztésének motorja, koordinátora, vezetője. Az alkalmazott nyelvészet területén kifejtett tudományos, tudományszervezői és vezetői munkássága egyaránt jelentős. A magyar alkalmazott nyelvészet ügyének aktív támogatója, munkássága jelentősen hozzájárult a magyar alkalmazott nyelvészet eredményeinek külföldi elismertségéhez, az utánpótlás-neveléshez.

A tudományszervezésben kifejtett munkája is jelentős. A nagyszámú nemzetközi projekt keretében és azon kívül is számos alkalmazott nyelvészeti (korpusznyelvészeti, számítógépes nyelvészeti, lexikográfiai) konferencia szervezője és társszervezője, programbizottságának és/vagy szervezőbizottságának tagja. Az MTA Alkalmazott Nyelvészeti Munkabizottságának elnöke 2005 és 2011 között, a munkabizottság által szervezett Alkalmazott Nyelvészeti Doktoranduszkonferenciák egyik szervezője és motorja, a 2007-ben indult Alkalmazott Nyelvészeti Doktoranduszkonferencia kiadványainak elindítója, jelenleg is sorozatszerkesztője (az első kötet: Váradi szerk., 2007). Két periódus után nem lehetett tovább a munkabizottság elnöke, de (fiatal munkatársak bevonásával) vállalta továbbra is a konferenciahelyszín biztosítását, a konferenciák szervezését, az Alkalmazott Nyelvészeti Doktoranduszkonferencia kiadványainak szerkesztését és megjelentetését. Az MTA Nyelvtudományi Intézete[1] saját honlapján biztosít online megjelenést a kötetek számára.

A nyelvtechnológia, a korpusznyelvészet, a lexikográfia, a szociolingvisztika területén jelentős munkásságán túl részt vett az ELTE Fordítástudományi Doktori Programjában, oktatóként a számítógépes alkalmazások és korpusznyelvészeti kutatások területén segítette a doktoranduszokat (lásd Klaudy, 2013). A Magyar Terminológia című folyóirat szerkesztőbizottságának tagja (2008–2013) és a Magyar Nyelv Terminológiai Tanácsának (MaTT) tagja 2013 óta.

2. Nyelvtechnológia, magyar nyelvi korpuszok, lexikográfia

Váradi Tamás 1997 óta a Korpusznyelvészeti Osztály, 2013 óta a Nyelvtechnológiai és Alkalmazott Nyelvészeti Osztály vezetője, tudományos főmunkatárs az MTA Nyelvtudományi Intézetében.

Nagyszámú nyelvészeti pályázati projekt elnyerése (vezetése, illetve részvétel) fűződik a nevéhez, például nagy nemzetközi projektek:

[1] 2019 szeptemberétől: Nyelvtudományi Intézet.

- 2011–2013: CESAR Central and South-East European Resources – EU CIP-ICT-PSP.2010.6.1;
- 2010–2012: iTranslate4.eu gépi fordító projekt – EU CIP-ICT-PSP.2009.5.1;
- 2007–2009: CLARIN Common Language Resources and Technology infrastructure – EU FP7 INFRA-2007-2.2.01.

Magyar projektek:

- 2005–2007: Magyar ontológia építése (EuroWordnet) – GVOP projekt;
- 2006–2008: Magyar Egységes Ontológia – NKFP projekt stb. (lásd www.nytud.hu).

Számos, a magyar nyelv feldolgozásának szempontjából nélkülözhetetlen munkálat vezetője, irányítója. Az irányításával létrejött magyar nyelvű és párhuzamos korpuszok, a nagy nemzetközi projektekben való részvétel megerősítette a magyarországi korpusznyelvészet és a lexikográfia elismertségét a széles körű nemzetközi és az európai kutatás-fejlesztésben. Néhány jelentősebb projektet emelek ki az alábbiakban (forrás: www.nytud.hu).

Az MNSz. (Magyar nemzeti szövegtár) munkálatai 1998 elején kezdődtek el a Magyar Tudományos Akadémia Nyelvtudományi Intézetének Korpusznyelvészeti Osztályán. Erre épült 2002-től a Kárpát-medencei Magyar korpusz, jelenleg pedig már az MNSz.[2] gigakorpusz munkálatai folynak a vezetésével. Az MNSz. a mai magyar írott köznyelv általános célú reprezentatív korpusza, online, ingyenesen elérhető magyar nyelvi korpusz (lásd pl. Oravecz és mtsai., 2014; Váradi és Oravecz, 2014).

Vezetésével kezdődött meg a helyesiras.mta.hu portál fejlesztése (lásd pl. Váradi és mtsai., 2014).

A kifejezetten lexikográfiai témájú projektek közül az EFNILEX keretében két- és többnyelvű úgynevezett protoszótárak félautomatizált létrehozását kísérelték meg, nyelvtechnológiai eszközökre és párhuzamos korpuszokra alapozva. Ennek keretében többszavas kifejezések kivonatolásával is kísérleteztek, ami a módszerek terminológiai alkalmazhatóságát is ígéretessé tette (lásd pl. Váradi és Héja, 2011; Váradi és Héja, 2012). Az Európai e-lexikográfiai hálózat (European Network of e-lexicography, ENEL) projekt keretében 2013 és 2017 között létrehoztak egy európai lexikográfiai portált[2] és ennek eredményeként European Lexicographic

[2] https://www.elexicography.eu

Infrastructure (ELEXIS)[3] néven elindítottak egy európai projektet. Ennek célja az európai szótári infrastruktúra felmérése és támogatása, az elektronikus, online szótárak és szótári applikációk készítésének támogatása, készítésük könnyebbé tétele oly módon, hogy harmonizálják az eredményeket és a szükségleteket, és mindenki által elérhető szabványokat és eszközöket fejlesztenek. Az egyetemi oktatók számára ezek az eszközök ingyenesen elérhetők a program honlapjáról.

Az általa vezetett kutatócsoport koordinálja a vezető hazai nyelv- és beszédtechnológiai kutatóhelyek stratégiai jelentőségű HunCLARIN kutatásiinfrastruktúra-hálózatának munkáját, melynek része az e-magyar.hu rendszer. A CLARIN egy olyan elosztott hálózati infrastruktúra szerte Európában, amely nem csak nyelvészek számára nyújt szolgáltatásokat, hanem minden olyan kutató számára, akik nyelvi erőforrásokat (korpuszokat, lexikai és egyéb adatbázisokat használnak) (lásd pl. Váradi és mtsai., 2018, Váradi és Jelencsik-Mátyus, 2020).

3. A terminológia mesterképzés nyelvtechnológiai, korpusznyelvészeti vonatkozásai

A terminológusok képzését, a terminológia intézményes, egyetemi szintű oktatását 2011-ben kezdtük meg a Károli Gáspár Református Egyetem Bölcsészettudományi Karának Magyar Nyelvtudományi Tanszékén.[4] Ehhez kapcsolódik az azóta is működő Terminológiai Kutatócsoport, amely a képzés terminológiai hátterét biztosította (lásd TERMIK). 2011 és 2016 között 6 évfolyam indult, 50 beiratkozott hallgatóval, akik közül 40-en szereztek egyetemi oklevelet a szakon. Az e képzésben részt vevő utolsó évfolyam 2018-ban végzett.[5]

A képzés felépítését, céljait, részleteit több tanulmányban tettük közzé (pl. Fóris, 2012; Fóris, 2013; Fóris és Bölcskei, 2019). A terminológia mesterszak létesítésének és indításának célja olyan szakemberek képzése volt, akik a terminológiatudomány vonatkozásában korszerű elméleti és módszertani ismeretekkel rendelkeznek, ismerik a terminológiai munka hazai és nemzetközi folyamatait és módszereit, továbbá akik jelentős szerepet tudnak vállalni a magyar nyelvvel és nyelvhasználattal kapcso-

[3] https://elex.is
[4] www.kre.hu/nyelveszet
[5] Az EMMI által 2016-ban kiadott felsőoktatási szakok jegyzéke nem tartalmazta a Terminológia mesterszakot (ennek okát nem sikerült kideríteni). Azóta a szak újraalapítására tett erőfeszítéseink nyomán a szak végigment a hivatalos engedélyeztetési folyamaton, minden engedélyt és támogatást megkapott, ennek ellenére egyelőre nem került bele az EMMI által kiadott felsőoktatási szakjegyzékbe.

latos terminológiai feladatok megoldásában. A képzést az MTA Nyelvtudományi Intézetén belül a Váradi Tamás által vezetett Nyelvtechnológiai és Alkalmazott Nyelvészeti Osztály tanácsokkal és a hallgatók számára nyújtott gyakornoki lehetőséggel támogatta. A képzés részét képezték a korpusznyelvészeti, adatbázis-kezelési, számítógépes nyelvészeti ismeretek és a fogalomalapú információkezelés oktatása. Annak érdekében, hogy megfelelő informatikai hátteret nyújthassunk a képzéshez, a Károli Gáspár Református Egyetem Bölcsészettudományi Karának Dózsa György úti épületében egy 20 fős számítógépes laboratóriumot alakítottunk ki. Ennek megtervezésében a matematika tárgy oktatója, Prőhle Péter (főállásban a BME egyetemi docense) volt nagy segítségünkre. E laboratórium nemcsak a megfelelő minőségű számítógépekkel lett felszerelve, hanem a szükséges szoftverekkel is, melyeknek egy része ingyenesen elérhető, más részét az egyetem megvásárolta, illetve egy részüknek az oktatási verzióját maguk a szoftvereket fejlesztő és forgalmazó cégek bocsátották a rendelkezésünkre (a terminológiakezelők közül pl. MemoQ és Trados).

A terminológia mesterszak keretében a szakmai törzstárgyak között szerepelt a *korpusznyelvészet* előadás és szeminárium, amely a négy féléves képzés második szemeszterében heti két előadásból és heti két szemináriumból állt. Ezen felül a kötelezően választható differenciált szakmai ismeretek keretében lehetőség volt további számítógépes nyelvészeti tárgyak felvételére a harmadik és negyedik szemeszterben. Ezek az alábbiak voltak: *adatbázisok kezelése* (2 óra előadás, 4 óra gyakorlat), *fogalomalapú információkezelés* (2 óra előadás, 2 óra gyakorlat), *számítógépes nyelvészet* (2 óra előadás, 2 óra gyakorlat). Az előadások és a szemináriumok tematikájának, anyagának kialakítása Váradi Tamás segítségével és támogatásával történt. E tárgyak felelőse és oktatója M. Pintér Tibor lett, aki az MTA Nyelvtudományi Intézetéből érkezett hozzánk kifejezetten e szakterület megerősítésére.

A *korpusznyelvészet előadás* célja a hallgatók bevezetése a korpusznyelvészet elméletébe és módszertanába. Elsajátítják, hogy mire jó a korpusznyelvészet és mire nem, mik a kísérleti célok, és melyek a korpusztervezés elvei és módszerei. A *szeminárium* célja az előadás tematikájához kapcsolódva megismerkedni különböző típusú korpuszokkal és azok használatával. A tematika részletesebben: Szövegek, fájlok, nyelvek a számítógépen. A korpusznyelvészet célja, módszerei. Rövid történeti áttekintés, ismerkedés jelentős korpuszokkal. A szöveg formai és tartalmi jegyeinek elkülönítése. A szövegannotálás célja, elve és technikája.

Szabványos jelölőnyelvek: HTML, SGML, XML. Tartalomjelölő szabványok: TEI, CES/XCES. Szövegfeldolgozó módszerek és eszközök. Reguláris kifejezések használata. A NooJ korpuszkezelő eszköz.

A hallgatók által elsajátítandó szakmai kompetenciák és megszerzendő ismeretek az alábbiak: megismerik a korpusznyelvészet elméletét és módszertanát, valamint alapfogalmait, és jártasságot szereznek bennük; jártasságot szereznek az új típusú korpuszokban és azok használatában, a szabványos jelölőnyelvek és vonatkozó szabványok ismeretében; felkészülnek a terminológiai munkák végzése során használt eszközök és módszerek alkalmazására, különösen a korpuszalapú, számítógéppel támogatott terminológiamenedzsment-eszközök használatára; képessé válnak új adatok bevitelére terminológiai adatbázisokba, terminológiai adatok exportálására más adatbázisokba, pl. egynyelvű vagy többnyelvű szótárakba.

Az *adatbázisok kezelése* című tárgy keretében az *előadás* célja az adatbázisokkal és az adatbázis-kezelőkkel kapcsolatos legfontosabb ismeretek összefoglalása. A kurzus témái: Az adatbázisokhoz kapcsolódó legfontosabb alapfogalmak, továbbá az adatbázisok kezelésének, használatának elsajátítása, valamint az adatbázisok különböző típusainak (pl. adatbank, adatbázis, tudásbázis, terminológiai adatbázis) megismerése. Az adatbázisok struktúrája, tartalma, az adatok bevitelének módszere. Adatbázis-kezelő szoftverek. Fogalom- és szóközpontú felfogás. A tantárgy épít a korpusznyelvészet tárgy keretében elsajátított ismeretekre – a szöveges adatbázis konvergál a korpusszal, és tulajdonképpen olyan annotációval van dolgunk, amelyből adatbázis építhető és viszont, az adatbázisból rekonstruálható a korpusz. A *gyakorlat* célja az adatbázisokkal való megismerkedés, a kooperáló partnerek gyakorlatából válogatott feladatokon keresztül. A terminológiai jellegű problémák megoldása adatbázisok segítségével, adatbevitel adatbázisokba, adatbázisok működtetése.

A hallgatók által elsajátítandó szakmai kompetenciák és megszerzendő ismeretek az alábbiak: megismerkednek a legújabb adatbázis-kezelőkkel és adatbázisokkal; megismerik a nyelvre vonatkozó empirikus adatok adatbázisban történő feldolgozási lehetőségeit, értelmezéseit; képessé válnak különböző nyelvi adatbázistípusokat munkájuk, kutatásuk során használni; képesek lesznek a magyar nyelvvel kapcsolatos számítógépes- és korpusznyelvészeti kutatásokba bekapcsolódni.

A *fogalomalapú információkezelés előadás* célja a hallgatók bevezetése a fogalomalapú információkezelés elméletébe. A kurzus témái: fogalom és fogalmi rendszerek; nyelvi és fogalmi rendszerek megfelelése; ontológia és számítógépes ontológia; tudásmenedzsment és tudásszervezés; tezauruszok, taxonómiák, fogalomalapú adatbázisok; fogalomalapú keresés és adatbevitel; hálóelmélet; a WordNet mint a hierarchikus lexikai viszonyok reprezentációja. A *gyakorlat* célja a fogalomalapú információkezelés módszereivel való megismerkedés, a kooperáló partnerek gyakorlatából válogatott feladatokon keresztül. A terminológiai jellegű problémák megoldása fogalomalapú adatbázisok segítségével, tezauruszok, ontológiák, és az ezeket kezelő szoftvereszközök megismerése.

A hallgatók által elsajátítandó szakmai kompetenciák és megszerzendő ismeretek: megismerkednek a fogalomalapú keresés és adatbevitel módszereivel; megismerik a fogalmakra vonatkozó adatok adatbázisban történő feldolgozási lehetőségeit, értelmezéseit; képessé válnak különböző fogalomalapú adatbázistípusokat munkájuk, kutatásuk során használni; képesek lesznek a magyar nyelvvel kapcsolatos számítógépes- és korpusznyelvészeti kutatásokba bekapcsolódni, és alkalmasak lesznek a fogalomalapú és a jelalapú szemlélet elkülönítésére.

A *számítógépes nyelvészet előadás* célja a számítógépes nyelvészet legfontosabb ismereteinek összefoglalása, a terminológiai munka során alapvető fontosságú, a hatékony terminológiai munkát segítő számítógépes eszközök megismertetése a hallgatókkal. Témák: a szövegszerkesztő, szövegelemző, szólistakészítő, formátumértelmező és formátumellenőrző programok; a modern nyelvleírásban nélkülözhetetlen eszközök, programok, a szövegalapú információs rendszerek formalizált szabványai: az SGML, a HTML és az XML, a web technológia, az annotációs technológia, a szabványok; statisztikai nyelvfeldolgozás, a terminuskivonatolás alapjai; a számítógépes nyelvészet lehetséges alkalmazási területei; a mesterséges intelligencia és a fordítói rendszerek kérdésköre. A *gyakorlat* célja elmélyíteni az előadáson megszerzett ismereteket, a kooperáló partnerek gyakorlatából válogatott feladatokon keresztül. A terminológiai jellegű problémák megoldása számítógépes nyelvészeti eszközök és módszerek segítségével, tezauruszok, számítógépes ontológiák és az ezeket kezelő szoftvereszközök megismerése.

A hallgatók által elsajátítandó szakmai kompetenciák és megszerzendő ismeretek: jártasságot szereznek a számítógépes nyelvészet elméletében és módszereiben; megismerik a terminológiai munkát segítő számítógépes eszközöket, programokat és szabványokat; képesek lesznek a

magyar nyelvvel kapcsolatos számítógépes- és korpusznyelvészeti kutatásokba bekapcsolódni, és alkalmasak lesznek a terminológia-menedzsment területén használt eszközök használatára.

E kurzusok keretében a kötelező és az ajánlott irodalom[6] angol és magyar nyelvű kézikönyvekből áll, ezek közül felsorolok néhány fontosabbat. Az angol nyelvűek: Biber, C., Reppen, R.: Corpus Linguistics. Cambridge University Press, Cambridge MA (1998); Fóris Á.: Change of paradigm in terminology: new models in KO. In: Gnoli, C., Mazzocchi, F. (eds.) Paradigms and conceptual systems in knowledge organization. Proceedings of the Eleventh International ISKO Conference 23–26 February 2010, Rome, Italy. pp. 57–63. Ergon Verlag, Würzburg (2010); McEnery, T., Wilson, A.: Corpus Linguistics. Edinburgh University Press, Edinburgh: (1996); Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Five Papers on WordNet. CSL Report 43. Cognitive Science Laboratory. Princeton University (1990); Sinclair, J.: Corpus, Concordance, Collocation. Oxford Univeristy Press, Oxford (1991); Stubbs, M.: Text and Corpus Analysis. Blackwell, Oxford (1996); Váradi T.: From Cards to Computer Files. Processing the Data of The Budapest Sociolinguistic Interview. Linguistics Institute, Hungarian Academy of Sciences, Budapest (1998), http://www.nytud.hu/buszi/wp3/index.html.

Az ajánlott magyar nyelvű szakirodalom: Demeczky J.: Terminológia a szoftveriparban. Magyar Terminológia 2/1, 189–204, (2008); Fóris Á.: A skálafüggetlen hálók nyelvészeti vonatkozásai. Alkalmazott Nyelvtudomány 7/1–2, 105–124 (2007); Kis B., Mohácsi-Gorove A.: A fordító számítógépe. SZAK Kiadó, Bicske (2008); Kiefer F. (szerk.) Strukturális magyar nyelvtan IV. A lexikon szerkezete. Akadémiai Kiadó, Budapest (2008); Pajzs J.: Számítógép és lexikográfia. MTA Nyelvtudományi Intézet, Budapest (1990); Prószéky G., Kis B.: Számítógéppel emberi nyelven. SZAK Kiadó, Bicske (1999); Prószéky G., Miháltz M.: Magyar WordNet: az első magyar lexikális szemantikai adatbázis. Magyar Terminológia 1/1, 43–58 (2008); Váradi T.: From Cards to Computer Files. Processing the Data of The Budapest Sociolinguistic Interview. Linguistics Institute, Hungarian Academy of Sciences, Budapest (1998), http://www.nytud.hu/buszi/wp3/index.html;

Ezen kívül online elérhető szótárakat, weboldalakat és szoftvereket használtunk az oktatásban, pl. Visuwords online graphical dictionary, http://www.visuwords.com; WordNet. Princeton University Cognitive

[6] Az oktatásban használatos fontosabb szakirodalmakat itt a szövegben felsoroltam, de nem írtam be a tanulmány végén a szakirodalomba, mert nem a szövegben hivatkozott munkákról van szó.

Science Laboratory, http://wordnet.princeton.edu; http://www.semantic-web.at; http://www.w3c.hu/.

M. Pintér Tibor, e tárgyak oktatója a *Digitális kompetenciák a felsőoktatásban* címmel írt tanulmányában (M. Pintér, 2019) összegezte a terminológia mesterszakos képzésben oktatott számítógépes nyelvészeti és infokommunikációs tartalmakra épülő tárgyak keretében szerzett tapasztalatait. Írásában hangsúlyozza, hogy célja volt, hogy a hallgatók elsajátítsák a legalapvetőbb információtechnológiai tudásanyagot és hogy megfelelő mértékben fejlessze a hallgatók digitális kompetenciáit:

> „A képzés folyamán elsajátítandó anyag (amely alapjában véve a digitális kompetenciák fejlesztésén alapult) elsősorban a gépi szöveg- és adatfeldolgozás változatos eszközeinek, valamint különféle módszereinek összességére fókuszált. Ennek megfelelően az általam elvárt tudást (ismeretanyag és kompetenciák) elsősorban az adatbányászatra építettem, fókuszba hozva az adatfeldolgozás, az adatbázisok és a statisztika alapjait, valamint a szövegközpontú megközelítés használatában kulcsfontosságú eljárásokat (ilyenek például a reguláris kifejezések ismerete és használata, az adatbázis-kezelés alapjainak mögöttes logikai megközelítése, alapvető, a statisztikában alkalmazott képletek, illetve a felhasznált adatok megközelítésének módjai – mindezt ama fontos kitétel mellett, hogy az adat önmagában, értelmezés és viszonyítás nélkül veszélyes és félrevezető)" (M. Pintér, 2019: 52–53).

M. Pintér az alapvető információ-feldolgozási műveletekre és szabványokra alapozta az oktatást, és ebben az említett tanulmányában részletesen ismerteti, hogy milyen programok, készségek és tudás elsajátíttatását látta szükségesnek a képzés során (M. Pintér, 2019: 54–55).

A terminológia mesterszakon végzett hallgatóink közül többen írtak szakdolgozatot korpuszok terminológiai szempontú vizsgálatáról, illetve felhasználásáról, pl. Erdős Róbert *Korpuszok a terminológiában. Az EUR-Lex jogszabálygyűjtemény vizsgálata terminológiai szempontból* címmel (Erdős, 2011; témavezető: Fóris Ágota és M. Pintér Tibor), Kovács (Dodé) Réka *Terminológia-menedzsment. Az ontológiai szemléletű terminológiai adatbázisok relációiról a WordNet, az EcoLexicon és az EOHS Term összevetése alapján* címmel (Kovács, 2012; témavezető: Fóris Ágota és Tamás Dóra Mária), Monostori Máté *Korpusztervezés és terminológia. Morfológiai elemzők terminológiai vizsgálata* címmel (Monostori, 2011; témavezető: Fóris Ágota és M. Pintér Tibor). Több

terminológia szakos hallgató kapott lehetőséget szakmai gyakorlata elvégzésére az MTA Nyelvtudományi Intézetében, Kovács (Dodé) Réka pedig az intézet Nyelvtechnológiai kutatócsoportjában helyezkedett el tudományos munkatársaként.

4. Összefoglalás

A magyar nyelvi adatbázisok és magyarországi korpusznyelvészet története egybefonódott Váradi Tamás személyével. A tudomány személyes dolog: adatokat meg lehet tanulni ugyan könyvekből, de tudományosan gondolkodni, a tudományos közösségbe beilleszkedni csak a közösség kapcsolatain keresztül lehetséges. Ő főként saját munkáján keresztül, az általa vezetett projektek keretében „nevelte ki" e nyelvészeti szakterület fiatal nemzedékét. Én magam elsősorban konferenciák szervezése, szerkesztések, lektorálások és az egyetemi képzés során kerültem szakmai kapcsolatba vele, és minden alkalommal tisztelettel figyeltem rendkívüli munkabírását, precíz időbeosztását, türelmét és könyörtelen logikáját, valamint mindezzel párosuló jó humorát és segítőkészségét. Ezt a most már másfél évtizede tartó szakmai együttműködést szeretném megköszönni Váradi Tamásnak ezzel a tanulmánnyal.

Bibliográfia

Erdős R.: Korpuszok a terminológiában. Az EUR-Lex jogszabálygyűjtemény vizsgálata terminológiai szempontból. MA-szakdolgozat, KRE BTK Magyar Nyelvtudományi Tanszék, Budapest (2011)

Fóris Á., Bölcskei A.: Szabványosítás, fordítás, terminológia. A szabványosítás a terminológia oktatásában. In: Fóris Á., Bölcskei A. (szerk.) A szabványosítás fordítási és terminológiai vonatkozásai. pp. 9−18. KRE, L'Harmattan, Budapest (2019)

Fóris Á.: Terminológusok képzése – a terminológia mesterképzés elindulása, Magyar Tudomány 173/8, 969–976 (2012) http://www.matud.iif.hu/2012/08/11.htm

Fóris Á.: A terminológia-oktatás mint a magyar nyelv, kultúra és identitás támogatása, Magyar Terminológia 6/2, 185–195 (2013)

Klaudy K.: A Fordítástudományi Doktori Program tíz éve. In: Klaudy K.: Fordítás és tolmácsolás a harmadik évezred elején: 40 éves az ELTE Fordító- és Tolmácsképző Tanszéke. pp. 19–31. ELTE Eötvös Kiadó, Budapest (2013)

Kovács (Dodé) R.: Terminológia-menedzsment. Az ontológiai szemléletű terminológiai adatbázisok relációiról a WordNet, az EcoLexicon és az EOHS Term összevetése alapján. MA-szakdolgozat, KRE BTK Magyar Nyelvtudományi Tanszék, Budapest (2011)

Monostori M.: Korpusztervezés és terminológia. Morfológiai elemzők terminológiai vizsgálata. MA-szakdolgozat, KRE BTK Magyar Nyelvtudományi Tanszék, Budapest (2011)

M. Pintér T.: Digitális kompetenciák a felsőoktatásban. Modern Nyelvoktatás 25/1, 47–58 (2019)

Oravecz, Cs., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014). pp. 1719–1723. European Language Resources Association (ELRA), Lisszabon (2014)
Váradi T.: NP Modification Structures in Parallel Corpora. In: Károly K., Fóris Á. (szerk.). New Trends in Translation Studies. In honour of Kinga Klaudy. pp. 191–206. Akadémiai Kiadó, Budapest. (2005)
Váradi T. (szerk.) I. Alkalmazott Nyelvészeti Doktorandusz Konferencia. Budapest, 2007.02.02, MTA Nyelvtudományi Intézet, Budapest, 2007. http://www.nytud.hu/alknyelvdok07/proceedings07/alknyelvdok07_online.pdf
Váradi, T., Héja, E.: Multilingual term extraction from parallel corpora. A methodology for the automatic extraction of verbal structures and their translation equivalents. Magyar Terminológia 4/2, 226–237 (2011)
Váradi, T., Héja, E.: EFNILEX Online Dictionaries. In: Stickel, G., Carrier, M. (eds.) Language Education in Creating a Multilingual Europe. pp. 165–180. Peter Lang, New York (2012)
Váradi T., Jelencsik-Mátyus K.: A CLARIN ERIC és HUNCLARIN bemutatása. 2020. december 15. https://www.youtube.com/watch?v=R-F0wVaX_wE (Letöltés: 2020. december 21.) Az előadás elhangzott az NKFI Hivatal által 2020. december 8-án szervezett „Magyar kutatók és kutatóhelyek kapcsolódásai nemzetközi kutatási infrastruktúrákhoz” című online rendezvényen
Váradi, T., Ludányi, Zs., Kovács, R.: Géppel segített helyesírás. A helyesírás.mta.hu készítéséről. Modern Nyelvoktatás 20/1–2, 43–58 (2014)
Váradi, T., Oravecz, Cs.: A Magyar nemzeti szövegtár egymilliárd szavas új változata. Magyar Tudomány 175/9, 1054–1061 (2014)
Váradi, T., Simon, E., Sass, B., Mittelholcz, I., Novák, A., Indig, B., Farkas, R., Vincze, V.: E-magyar – A Digital Language Processing System. In: Calzolari, N., Choukri, K., Cieri, Ch., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (eds.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). pp. 1307–1312. European Language Resources Association (ELRA), Paris (2018)

Források

ELEXIS (European Lexicographic Infrastructure), https://elex.is/ (Letöltés: 2020. december 22.)
ENEL (European Network of e-lexicography), https://www.elexicography.eu/ (Letöltés: 2020. december 22.)
KRE BTK MNYIKI Magyar Nyelvtudományi Tanszék, www.kre.hu/nyelveszet (Letöltés: 2020. december 22.)
Nyelvtudományi Intézet, www.nytud.hu (Letöltés: 2020. december 22.)
TERMIK (Terminológiai Kutatócsoport), https://btk.kre.hu/index.php/2015-12-05-09-31-20/kari-kutatocsoportok/824-terminologiai-kutatocsoport-termik.html (Letöltés: 2020. december 22.)

Az EFNILEX és egy fiatal kutató. Hat év magyar szóbeágyazásokkal

Makrai Márton[1,2]

[1] BME VIK Távközlési és Médiainformatikai Tanszék
[2] MTA TTK Kognitív Idegtudományi és Pszichológiai Intézet
makrai.marton@ttk.hu

1. Európai szótárak egynyelvű korpuszból

Az EFNILEX projekt azt szándékozott felderíteni, hogy a gépi fordítás eszközei hogyan járulnak hozzá szótárak előállításához „közepes" európai nyelveken, vagyis az EU kevesebb nyelvtechnológiai erőforrással rendelkező hivatalos nyelvein. Héja Enikőtől vettem át a stafétát 2014-ben, aki – ahogy ebben a kötetben is írja – párhuzamos korpuszokból készített szótárakat.

Ezekben az években volt a nyelvtechnológia neurális forradalmának első, sekélyebb hulláma, az előtanított (de nem kontextualizált, nem igazán mély) szóbeágyazásoké. A *beágyazás* szó arra utal, hogy a szimbolikus készlet elemei (esetünkben: szókincs) a neurális hálókban valós vektorként vannak reprezentálva. A 2013–14-es hullám első cikkei több mindenre rámutattak: Ezek a módszerek a korábbinál sokkal hatékonyabban állítanak elő egynyelvű szemigigakorpuszból – amilyen magyarra nem utolsó sorban az MNSz.[2] (Oravecz és mtsai., 2014) – jó minőségű szóreprezentációkat: a jelentésben vagy morfoszintaktikailag hasonló szóalakokat egymáshoz közel helyezik el egy pár száz dimenziós euklideszi térben (Mikolov és mtsai., 2013a, 2013b). Továbbá a lexikai szemantika régi álmához, a szótári felbontáshoz is közelebb vittek: Mikolov és mtsai. (2013c) híres és sokat vitatott (Levy és mtsai., 2015; Linzen, 2016) példájával a *királynő* fogalma egy *királyi* és egy *női* elemből áll. Végül – és az EFNILEX e szakasza szempontjából első sorban – a különböző nyelvek egynyelvű korpuszaiból tanított modellek között olyan hasonlóság áll fönn, amely lehetővé teszi az úgynevezett lineáris fordítást, vagyis hogy egy néhány ezer szavas magszótárból föltanítsunk egy lineáris leképezést a két nyelv szavainak vektortere között, amellyel a forrásnyelvi w_s szót a célnyelvi térbe képezve egy olyan vektort kapunk, amelyhez legközelebbi célnyelvi szóvektor formájában megtaláljuk w_s célnyelvi megfelelőjét.

DOI: https://doi.org/10.18135/VT70.12

Enikő bátorított minket, hogy alkalmazzuk a lineáris fordítás módszerét az EFNILEX-ben. A szóvektorok kiértékelésének egyik legnépszerűbb módszerét az analógiás kérdések jelentik, pl. *férfi : nő :: király : ?,* a várt válasz a *királynő.* Elkészítettük és nyíltan közreadtuk (Makrai, 2014) az egyik fő (angol) teszthalmaz magyar megfelelőjét.

Ezeknek a sekély neurális háló segítségével előállított modelleknek a mai mély előtanított neurális nyelvmodellek kontextualizált szóbeágyazásaival szemben (lásd az utolsó szakaszt) az volt a hátrányuk, hogy egy szóalakot, legyen az *poliszém* vagy akár *homonim*, egyetlen vektor reprezentált, a különböző jelentések a legjobb esetben szuperponálódtak, vagy a ritkább jelentés elveszett, esetleg káros módon keveredtek. Prószéky Gábor példájával élve: tudomásul kellett venni, hogy a *daru* egy olyan entitás, amelyik olykor fészket rak, máskor betondarabokat emelget. 2015-től az EFNILEX-ben, majd fiatal kutatóként ezen a hiányosságon igyekeztünk-igyekeztem javítani kétféleképpen.

2. Szótári háromszögek egyértelműsítése

A gépi szófordítás (avagy szótárindukció) egyik bevett eszköze az úgynevezett *háromszögelés (triangulation)* vagy *sarokkőmódszer (pivot-based method).* Abból, hogy a cseh *zvíře* angol fordítása *animal*, az *animal* magyar fordítása pedig *állat*, arra lehet következtetni, hogy a *zvíře* magyarul *állat*. Ebbe a módszerbe több úton is zajt hoz a többértelműség. A középső nyelv homonímiái hamis háromszögeket vezetnek be (német *was* – magyar *mi* – angol *we*). A vektoros módszer viszont csak a forrás- és a célnyelv többértelműségeire érzékeny, így a két hiba kompenzálja egymást. Makrai (2016) hamis háromszögeket szűrt ki szóvektorok segítségével a német–magyar nyelvpáron. Megmutattuk, hogy a lineáris leképezésből kapott pontszámok simább mértékét adják a fordítások jóságának, mint ha csak megszámoljuk, hogy hány nyelven keresztül háromszögelhető az adott szópár. A nyíltan közreadott, megbízhatósági pontszámokkal ellátott német–magyar erőforrás tudomásunk szerint a legnagyobb szabad elérésű szólista volt akkor.

3. Egy fiatal kutató és a túl finom jelentéskészlet

2015-től 2018-ig az intézet fiatal kutatója voltam Tamás vezetésével. Nagyon hálás vagyok, hogy 3 éven át teljes állásban kutathattam a témámat, és a legjobb konferenciákon publikálhattam az eredményeket. Mint

mondtuk, a szóvektorok a szokásos esetben egy-egy szóalakhoz tartoznak, így a többértelmű szavak vektora rosszabb minőségű. Ezt a problémát hivatottak megoldani a többjelentésű szómodellek *(multi-sense word embedding, MSE)*, amelyek a szóalakok különféle jelentéseit különböző vektorokkal ábrázolják. Ebben a paradigmában annak a megállapítása is a felügyeletlen modell feladata, hogy mely szavak többértelműek, és azoknak hány jelentése van. Az alkalmazásban legjobbnak bizonyuló modellek vektorai közül azonban sok nem felel meg a motiváló várakozásoknak: jobb esetben olyan jelentések között tesznek különbséget, melyeket intuitíve ugyanazon jelentés különböző kontextusokban való használatának tekintenénk, vagy akár puszta zajt képviselnek.

Ezért a szerzőtársaimmal (Borbély és mtsai., 2016) két új módszert javasoltunk az MSEk szemantikai szemcsésségének mérésére. Az egyik egynyelvű szótárakat használ, a másik pedig azon az elven alapszik, hogy egy szó akkor többértelmű, ha a feltételezett jelentések más nyelvre való fordítása különböző. Az utóbbit Makrai és Lipp (2019) bontotta ki két pontosságmértéket formalizálva. Az egyik bünteti a duplumokat, a másik pedig azért van, hogy a vektorok ne mossanak össze jelentéseket. A kísérletek igazolták, hogy a két mérték között csereviszony van: minél specifikusabb egy vektor, annál könnyebb lefordítani, csak persze ha túl specifikus, akkor egybeeshetnek a fordítások. Tehát a két mérték számszerűsíti, hogy egy többjelentésű szóbeágyazás mennyire jól ragadja meg a lexikai struktúrát (Borbély és mtsai., 2016; Makrai és Lipp, 2019).

A kutatás egy másik ágában Berend Gáborral hipernimákat (a fölérendelt fogalmat, pl. hogy a kutya egy állat) nyertünk ki szóvektorokból. Ritka szóreprezentációkon alapuló módszerünkkel megnyertünk több kategóriát a szakma évente megrendezésre kerülő legrangosabb versenyének (SemEval) egyik feladatában. Abban az évben New Orleansben volt a SemEval, így nem tudtam volna prezentálni a posztert, ha nincs az a nagyon stabil és bőséges anyagi keret, amit Tamás biztosított.

Bár nem tartozott projektbe, témája miatt kedves volt Tamásnak egy olyan cikk, amely egy általam bírált szakdolgozatból született: szónál kisebb elemek beágyazásán alapuló magyar nyelvmodelleket hasonlított össze a hallgató (Döbrössy és mtsai., 2019).

4. Evezz a mélyre

2018-ban újabb, mélyebb hullámot vetett a nyelvtechnológia (NLP) neurális forradalma. Mély neurális hálóval való tanulás alatt azt értjük,

hogy a gépi tanulás eredménye egy olyan számítási modell, amely rétegekből áll, és az input rétegtől rejtett rétegeken át az output réteg felé haladva egyre magasabb szintű jellemzőket számít ki. A mélytanulás először a beszédtechnológiában (Dahl és mtsai., 2011) és a gépi látásban (Krizhevsky és Sutskever, 2012) hozott áttörtést. 2018-ban az NLP-ben is elérkezett az, amit Sebastian Ruder *ImageNet pillanat*nak[1] nevez.

> „A gépi látás (*computer vision*, CV) kutatóközössége évek óta tanít fel teljes modelleket alacsony és magas szintű jellemzők előtanításával. Leggyakrabban ez úgy történik, hogy a nagy ImageNet adatkészlet képeinek osztályozását tanítják meg. Az ULMFiT, az ELMo és az OpenAI transzformer most elhozta a nyelv ImageNet-jét, vagyis egy olyan feladatot, amely lehetővé teszi a modellek számára, hogy a nyelv magasabb szintű aspektusait is megtanulják a modellek, hasonlóan ahhoz, ahogy az ImageNet lehetővé tette olyan CV-modellek feledzését, amelyek a képek általános célú jellemzőit tanulják meg."

Az utóbbi két évről kiváló áttekintést adnak Qiu és mtsai. (2020). A számítógépes nyelvész számára különösen érdekes a modellek nyelvészeti tudásának letapogatására irányuló kutatás, amit Rogers és mtsai. (2020) foglalnak össze. 2020-ban elindult egy magyar mély nyelvmodellek létrehozására, kiértékelésére, és nyelvészeti tartalmának felderítésére irányuló projekt is (HILBERT, Feldmann és mtsai., 2021).

Bibliográfia

Borbély, G., Makrai, M., Nemeskey, D. M., Kornai, A.: Evaluating multi-sense embeddings for semantic resolution monolingually and in word translation. In: Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP. pp. 83–89. Association for Computational Linguistics, Berlin (2016), http://www.aclweb.org/anthology/W16-2515

Dahl, G. E., Yu, D., Deng, L., Acero, A.: Large vocabulary continuous speech recognition with context-dependent dbn-hmms. In: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. pp. 4688–4691. IEEE (2011)

Döbrössy, B., Makrai, M., Tarján, B., Szaszák, G.: Investigating sub-word embedding strategies for the morphologically rich and free phrase-order Hungarian. In: Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019). pp. 187–193. Association for Computational Linguistics, Florence, Italy (2019) https://www.aclweb.org/anthology/W19-4321

[1] https://ruder.io/nlp-imagenet/

Feldmann, Á., Váradi, T., Hajdu, R., Indig, B., Sass, B., Makrai, M., Mittelholcz, I., Halász, D., Zijian, G. Y.: HILBERT, magyar nyelvű bert-large modell tanítása felhő környezetben. In: Berend G., Gosztolya G., Vincze V. (szerk.) XVII. Magyar Számítgépes Nyelvészeti Konferencia, pp. 29–36. Szegedi Tudományegyetem TTIK, Informatikai Intézet, Szeged (2021) MSZNY (2021)

Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25 (NIPS 2012) (2012)

Levy, O., Remus, S., Biemann, C., Dagan, I.: Do supervised distributional methods really learn lexical inference relations? In: Mihalcea, R., Chai, J., Sarkar, A. (eds.) Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 970–976. Association for Computational Linguistics (2015)

Linzen, T.: Issues in evaluating semantic spaces using word analogies. In: RepEval (2016)

Makrai, M.: Deep cases in the 4lang concept lexicon. In: Tanács, A., Varga, V., Vincze, V. (szerk.) X. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2014). pp. 50–57. Szegedi Tudományegyetem, Szeged (2014)

Makrai, M.: Filtering wiktionary triangles by linear mapping between distributed models. In: LREC (2016)

Makrai, M., Lipp, V.: Do multi-sense word embeddings learn more senses? In: Gyuris, B., Mády, K., Recski, G. (eds.) K + K = 120 Workshop Dedicated to László Kálmán and András Kornai on the occasion of their 60th birthdays. pp. 385–398. (2019) MTA Research Institute for Linguistics, Budapest,

Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation (2013a), arXiv preprint arXiv:1309.4168

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. (eds.) Advances in Neural Information Processing Systems 26, pp. 3111–3119. Curran Associates, Inc. (2013b), https://bit.ly/39HikH8

Mikolov, T., Yih, W.t., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013). pp. 746–751. Association for Computational Linguistics, Atlanta, Georgia (2013c)

Oravecz, C., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In: Calzolari, N. et al. (eds.) Proceedings of Ninth International Conference on Language Resources and Evaluation (LREC 2014). pp. 1719–1723. Reykjavik. ELRA. (2014) http://www.aclweb.org/anthology/L14-1536

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X.: Pretrained models for natural language processing: A survey. arXiv preprint arXiv:2003.08271 (2020)

Rogers, A., Kovaleva, O., Rumshisky, A.: A primer in bertology: What we know about how bert works. arXiv preprint arXiv:2002.12327 (2020)

A CLARIN és a HunCLARIN

Jelencsik-Mátyus Kinga[1]

[1] Nyelvtudományi Intézet
matyus.kinga@gmail.com

1. Bevezetés

Bár a magyar nyelvet az Európai Unió az erőforrásokkal kevésbé ellátott nyelvek közt tartja számon, a nyelvtechnológiával foglalkozó kutatóközpontokban, egyetemeken, archívumokban mégis jó néhány nyelvi korpusz megtalálható a kisebb speciális korpuszoktól (mint például a BioScope korpusz)[1], a több millió szóból álló írott egynyelvű korpuszokon át (lásd például a Magyar nemzeti szövegtár 2. változatát)[2], az írott és beszélt többnyelvű vegyes korpuszokig (ilyen például az Uráli adatbázis)[3]. Az adatgyűjtemények mellett több magyar kutatóközpontban és egyetemen foglalkoznak nyelvtechnológiai elemzők létrehozásával (lásd például az e-magyar elemzőrendszert)[4]. Ezek a nyelvtechnológiai eszközök jelentősen megkönnyítik nagyobb mennyiségű nyelvi adat feldolgozását a (főként) bölcsészet- és társadalomtudományok kutatásaiban. De még hiányzik egy láncszem: Honnan fognak tudomást szerezni a nem nyelvész kutatók ezekről a lehetőségekről? Kitől kapnak szakmai segítséget a nagy nyelvi adatbázisok feldolgozásához? Sőt, akár a nyelvész kutatók is hogyan fogják megtudni, hogy esetleg más nyelveken vannake már bevált módszerek egy felmerülő probléma megoldására?

2006-ban több mint 20 európai ország nyelvtechnológiával foglalkozó szakemberének részvételével, Váradi Tamás meghívására az MTA Nyelvtudományi Intézetben tartották a CLARIN előkészítő találkozóját. Ez a szervezet épp a fent bemutatott hiányzó láncszem létrehozását tűzte ki céljául.

[1] https://rgai.inf.u-szeged.hu/node/105
[2] http://mnsz.nytud.hu/
[3] http://www.nytud.hu/oszt/elmnyelv/urali/adatbazisok.html
[4] https://e-magyar.hu/hu/

DOI: https://doi.org/10.18135/VT70.13

2. A CLARIN

2.1. A CLARIN célja

A CLARIN (Common Language Resources and Technology Infrastructure) egy európai kutatásiinfrastruktúra-hálózat, amely a digitális nyelvi adatbázisokat és nyelvi feldolgozóeszközöket elérhetővé teszi a bölcsészettudományok és a társadalomtudományok kutatói számára. Kiindulópontja az az elképzelés, hogy az európai és azon túli nyelvek digitális nyelvi erőforrásait egyetlen internetes portálon összefogva egyszerűen hozzáférhetővé tegye. A CLARIN lényegében nem más, mint egy diffúz infrastruktúra, tagintézményekkel (egyetemek, kutatóintézetek) szerte Európában, amelyek szigorú elvárások alapján elnyerhetik a Centre B (K, C, stb.) státuszt.

2.2. Előkészítő szakasz

Két hónappal az MTA Nyelvtudományi Intézetben tartott előkészítő találkozó után benyújtották a CLARIN előterjesztését az Európai Bizottsághoz, majd 2008-ban elkezdődhetett az előkészítő szakasz 22 ország közreműködésével.

Az előkészítő szakasz 36 hónapja alatt megteremtették a megosztott infrastruktúra alapjait. Elsőként kidolgozták az infrastruktúra létrehozásának és működtetésének pénzügyi és irányítási alapelveit, amelyet később az összes részt vevő ország aláírt. A második, kihívást jelentő feladat az addig példa nélküli technikai háttér kialakítása volt, amely lehetővé teszi az összes felmerülő nyelv adatbázisaihoz és nyelvfeldolgozó eszközeihez való egyszerű, egy elérési ponton keresztüli hozzáférést. Harmadikként az infrastruktúra tényleges kialakításához és működésének teszteléséhez a prototípust fel kellett tölteni nyelvi erőforrásokkal minden részt vevő nyelvből. Ebben egyrészt felhasználták a már meglévő korpuszokat és eszközöket, másrészt rávilágítottak arra, hogy számos nyelvben alapvető nyelvi erőforrások is hiányoznak. Ezek létrehozása már a következő szakasz egyik célja lesz. Az előkészítő szakasz negyedik, legfontosabb feladata a felhasználók feltérképezése. Megvizsgálták, mely nyelvtechnológiai folyamatokat használják a leginkább a bölcsész- és társadalomtudományokban. Több kutatásban letesztelték az infrastruktúra használhatóságát. Kiemelten fontosnak tartották, hogy együttműködéseket alakítsanak ki bölcsészek és nyelvtechnológusok között (Váradi és mtsai., 2008).

A szakasz zárótalálkozóját szintén az Intézetben tartották 2011 júniusában.

2.3. Építő szakasz

A CLARIN ERIC (European Research Infrastructure Consortium) 2012-ben jött létre az Európai Bizottság döntése alapján, azzal a céllal, hogy létrehozza és fenntartsa az infrastruktúrát, amely támogatja a nyelvi adatok és eszközök megosztását, használatát és fenntarthatóságát főként a bölcsészet- és társadalomtudományok számára. A CLARIN ERIC-nek tagja lehet ország vagy kormányközi szervezet. Magyarország, bár a kezdetektől jelen volt a folyamatokban, csak 2016. augusztus 1-jén csatlakozott hivatalosan is a konzorciumhoz. A CLARIN-nak jelenleg 21 tagja és 3 megfigyelő státuszú országa van. Az egyes országokon belül a tagok (jellemzően kutatóintézetek, egyetemek, könyvtárak, archívumok) létrehoznak egy nemzeti konzorciumot. A CLARIN tehát egy szétszórt infrastruktúra szerte Európában, ahol a tagok nyelvi korpuszokat, digitális nyelvfeldolgozó eszközöket, valamint szakmai segítséget nyújtanak a nyelvi anyagokkal dolgozó kutatóknak.

Az infrastruktúra gerincét a központok alkotják. Központ lehet minden olyan intézmény vagy nemzeti konzorcium, amely megfelel a szigorú elvárásoknak, és végigmegy az engedélyeztetés folyamatán. A legfontosabb központtípus a B, a szolgáltatást nyújtó központ. Ezek alkotják a CLARIN magját. Ezek a központok olyan szolgáltatásokat nyújtanak, amelyek többek közt hozzáférést biztosítanak az általuk tárolt nyelvi korpuszokhoz, és az általuk kifejlesztett eszközök folyamatosan elérhetőek valamely CLARIN-nak megfelelő felületen.

A K központok tudásközpontok, amelyek szakmai segítséget nyújtanak a kutatónak ahhoz, hogy használni tudják a CLARIN nyújtotta szolgáltatásokat. Az egyes K központok eltérő területeken segítik a kutatókat. A C központok metaadatokat szolgáltatnak folyamatosan elérhető módon. Az E központok külső központok, amelyek a CLARIN-hoz kapcsolódó szolgáltatásokat nyújtanak, de nem a CLARIN tagjai. A CLARIN jelenlegi központjai láthatóak az 1. képen.

1. kép. A CLARIN központjai.[5]

2.4. Üzemeltetési szakasz

Ma körülbelül 20 B, és számos más típusú központ van a CLARIN-ban, számuk folyamatosan növekszik, a szervezet tehát a különböző központok hálózataként működik. A gondos előkészítés után a több éves működés alapján látható, hogy a CLARIN egyszerű és fenntartható hozzáférést nyújt a digitális nyelvi adatokhoz (írott, beszélt vagy multimodális) a bölcsészet- és társadalomtudományok kutatóinak. Fejlett eszközöket biztosít a nyelvi adatok kutatására, elemzésére. Lehetőséget nyújt a nyelvi korpuszok és eszközök kombinálására, összehasonlítására, valamint szakmai segítséget kínál mindezek használatához (Jong és mtsai., 2018). Technikai háttér tekintetében nyelviadat-repozitóriumok, szolgáltató központok és tudásközpontok állnak a részt vevő országok kutatói szolgálatában, egy egyszerű single sign-on eléréssel. Elmondható tehát, hogy az adatok és eszközök interoperabilitása megvalósult (Hinrichs és Krauwer, 2014).

A CLARIN ma számos országban tökéletesen működik. A meglévő korpuszok és eszközök fejlesztéséhez segítséget nyújtanak, az újonnan jelentkező országokban pedig segítik a rendszer kiépítését.

[5] A kép forrása: https://www.clarin.eu/content/overview-clarin-centres

3. A HunCLARIN

A HunCLARIN a vezető hazai nyelv- és beszédtechnológiai kutatásfejlesztést végző tudásközpontok stratégiai jelentőségű kutatásiinfrastruktúra-hálózata (SKI).

A kutatások bázisát képező nyelvi erőforrásokat és eszközöket tartalmaz. A megosztott virtuális hálózat 2010-ben, majd 2015-ben ismét SKI minősítést kapott. A HunCLARIN-hoz eddig 8 partner csatlakozott:[6] Nyelvtudományi Intézet (mint a HunCLARIN központja), BME Média Oktató- és Kutatóközpont, BME Távközlési és Médiainformatikai Tanszék, Szegedi Tudományegyetem, Debreceni Egyetem, Pázmány Péter Katolikus Egyetem, Morphologic Kft., valamint a Számítástechnikai és Automatizálási Kutatóintézet.

Az ezekben a központokban létrehozott jelenleg több mint 40 tag számos általános és speciális szövegkorpuszt, különféle nyelvi feldolgozó eszközöket, elemzőket, adatbázisokat, ontológiákat ölel fel.[7,8] A hálózat koordinátora és kapcsolattartója: Váradi Tamás.

A HunCLARIN legfontosabb célja a tudományos kutatás támogatása a nyelvtechnológia, a nyelvi erőforrások könnyű elérhetővé tételével. Ennek alapfeltétele egy olyan internetes felület, valamint az annak hátterében álló technikai infrastruktúra létrehozása, amelyen keresztül (a regisztrált kutatók számára) a csoportban található összes KI egyszerűen elérhető, valamint az eszközök egymással és a CLARIN más nyelveken megvalósuló alkalmazásaival összevethető. Ezzel lényegesen egyszerűbbé válik a magyar nyelv- és beszédtechnológia bekapcsolása a magas szinten folyó európai munkálatokba, hiszen a CLARIN számos más európai tagjánál (és azok között) a nyelvtechnológiai eszközök és erőforrások interoperabilitása már megvalósult.

A HunCLARIN tagjai számos jelentős hazai és nemzetközi projektben vettek részt. Ilyen például az uráli–orosz kontaktushatás kutatását is lehetővé tevő többnyelvű Uráli adatbázis, amely írott és beszélt nyelvi szövegeket is tartalmaz udmurt, tundrai nyenyec, szinjai és szurguti hanti nyelven.

Ahogy az 1. képen látszik, Magyarországra még nincs központ jelölve, de a HunCLARIN célja a B központ státusz elérése.

---

[6] http://clarin.hu/content/hunclarin-tagjai
[7] http://clarin.hu/content/korpuszok
[8] http://clarin.hu/content/nyelvtechnol%C3%B3giai-eszk%C3%B6z%C3%B6k

4. A felhasználók bevonása

A CLARIN, és vele összhangban a HunCLARIN is nagy hangsúlyt fektet a felhasználók, illetve a leendő felhasználók bevonására, tájékoztatására. Konzorciumon belül, tehát a magyarországi tagok közt, valamint nemzetközi szinten is évente számos alkalommal rendeznek előadásokat, workshopokat és webináriumokat. Ezek során nagy hangsúlyt fektetnek arra, hogy a résztvevőknek lehetőségük legyen kötetlen módon információkat szerezniük.

A CLARIN, illetve a HunCLARIN bemutatásának, valamint a más kutatóközösségekkel való kapcsolatépítésnek egyik nagyon hatékony módja a roadshow. Ezt bizonyítja az eddig megrendezésre került 3 rendezvény is Szegeden, Debrecenben, illetve Pécsett.

A roadshow lényege, hogy házhoz viszi a nyelvtechnológiát oda, ahol a bölcsész és társadalomtudományi kutatások zajlanak, vagyis az egyetemekre. Ezeknek az eseményeknek a szerkezete mindig úgy épül fel, hogy a nap kezdetén a HunCLARIN központból érkező nyelvtechnológusok röviden ismertetik a HunCLARIN, illetve a CLARIN célkitűzéseit, felépítését, működését, majd bemutatják, milyen korpuszokat és nyelvfeldolgozó eszközöket nyújthatnak a kutatók számára. A második részben a helyi bölcsészettudományi műhelyekben zajló munkákba kaphatunk betekintést, amelyekben nyelvtechnológiai eszközöket is igénybe vettek a nyelvi adatok elemzéséhez. Mindkét részben nagy hangsúlyt fektettek a közönség és az előadók közti párbeszédre.

Bibliográfia

Hinrichs, E., Krauwer, S.: The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 1525–1231. European Language Resources Association (ELRA), Reykjavik, Iceland (2014)

Jong, F. de, Maegaard, B., De Smedt, K., Fišer, D., Van Uytvanck, D.: CLARIN: Towards FAIR and Responsible Data Science Using Language Resources. In: Calzolari, N. et al. (eds) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), May 2018. pp. 3259–3264. European Language Resources Association (2018)

Váradi, T., Wittenburg, P., Krauwer, S., Wynne, M., Koskenniemi, K.: CLARIN: Common Language Resources and Technology Infrastructure. In: Calzolari, N. et al. (eds) Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). pp. 1244–1248. European Language Resources Association (2008)

Tamás Váradi and the International Lexicography

Lipp Veronika[1]

[1] Nyelvtudományi Intézet
lipp.veronika@nytud.hu

1. European Network of e-Lexicography (2013–2017)

In the last 10 years a clear need has arisen for a common approach to e-lexicography that forms the basis for a new type of lexicography fully embracing the pan-European nature of many of the vocabularies of languages spoken in Europe.

The aim of the COST European Network of e-Lexicography (ENeL, 2013–2017) was to increase, co-ordinate and harmonize European research in the field of e-lexicography and to make authoritative information on the languages of Europe easily accessible.

Tamás Váradi joined the project in 2013 as the leader of the Department of Language Technology and Applied Linguistics of the Research Institute for Linguistics. He became one of the most active members of the Management Committee of the network.

The aim of the network was to make lexical knowledge of small and large languages available in a European dictionary portal. The portal was finalized in 2017 and it serves as the central reference point for all dictionary users who look for reliable, authoritative dictionary information on the languages of Europe and their histories.[1]

The network enabled cooperation and the exchange of resources, technologies and experience in e-lexicography and provided support for dictionaries which were not online yet.

There were numerous meetings in which we discussed establishing standards for innovative e-dictionaries that fully utilize the possibilities of the digital medium and tried to establish new ways of representing the common heritage of European languages by developing shared editorial practices and by interconnecting already existing information.

Tamás Váradi was involved mainly in two working groups' tasks: one was engaged in the integrated interface for European dictionary content,

[1] http://www.dictionaryportal.eu/hu/

DOI: https://doi.org/10.18135/VT70.14

whereas the other investigated how the pan-European nature of the vocabularies of the languages of Europe could be represented in monolingual dictionaries and within the European dictionary portal.

The first working group investigated how authoritative dictionary information on the languages of Europe can be made accessible to both the general and academic public. A European dictionary portal was set up, which gives information on scholarly dictionaries of the languages of Europe and provides access to these dictionaries. It investigated the possibilities of interlinking the contents of European dictionaries, explored user requirements with respect to the presentation of dictionary content, and mapped the possible involvement of users in the creation of dictionary content.

The other working group investigated how the pan-European nature of the vocabularies of the languages of Europe could be represented in monolingual dictionaries and within the European dictionary portal.

Tamás organized the 6th action meeting in Budapest with great success in February, 2017.

2. European Lexicographic Infrastructure

Another lexicographical project called ELEXIS started in 2018. It was a continuation of the ENel project, and it aimed at developing lexicographical tools. The lexicographical landscape in Europe is currently rather heterogeneous. There are standalone lexicographical resources, which are typically encoded in incompatible data formats due to the isolation of efforts, disabling the reuse of data in natural language processing, linked open data and the Semantic Web, or in digital humanities. ELEXIS will introduce common standards, develop conversion tools, and most importantly, it will interconnect the existing resources so that they can be used to develop new modern data which can be used in ways that new digital technologies require.

The main aim of the ELEXIS project is to introduce common standards, develop conversion tools, and most importantly, it will interconnect the existing resources so that they can be used to develop new modern data which can be used in ways that new digital technologies need.

The Research Institute for Linguistics is actively involved in this project under Tamás Váradi's leadership. Aligning senses across resources and languages is a challenging task with beneficial applications in the field of natural language processing and electronic lexicography. In this project the goal was to align word senses in monolingual dictionaries.

The alignment is carried out at sense-level for various resources in Hungarian and in 14 other languages (Ahmadi et al., 2020). Another task is to create parallel sense-annotated datasets in many European languages to be used for evaluation purposes in word-sense disambiguation.

The main merit of Tamás Váradi is that he has placed Hungarian lexicographical research into the international context. He has established extensive relationships with international lexicographical institutions and prominent lexicographers. I am very grateful for having been able to work with him over the past few years and I hope it will continue long into the future.

Dear Tamás, I wish you a happy birthday and much success in your further career.

References

Ahmadi, S., McCrae, J P., Nimb, S., Khan, F., Monachini, M., Pedersen, B. S. Declerck, T., Krek, S., Lipp, V., Váradi, T. et al.: A Multilingual Evaluation Dataset for Monolingual Word Sense Alignment. In: Calzolari, N. et al. (eds.) Proceedings of the 12th Language Resources and Evaluation Conference. pp 3232–3242. ELRA (2020)

MORENA – a project never realized (so far)

Marko Tadić[1]

[1]University of Zagreb, Faculty of Humanities and Social Sciences
marko.tadic@ffzg.hr

1. Introduction

After the successful completion of the ICT-PCP project CESAR within the META-NET initiative in January 2013, some of the partners under the coordination of Tamás Váradi wanted to prolong their fruitful collaboration and started seeking funding opportunities. At that time it seemed that the funding for LT would be scarce and that opportunities similar to the existing ones in earlier times would not be easily available. Therefore, we engaged in finding opportunities beyond our traditional funding sources and programmes that we had known before.

First, we tried our hands at the Scientific support to Danube Strategy. After two meetings including presentations of the CESAR project results with Tamás Váradi, Marko Tadić, and Dan Tufiș1 in attendance in Ispra in March and July 2013, we genuinely believed that we would be able to find funding for the jointly proposed project(s). However, this turned out to be true more or less exclusively for JRC in Ispra, whereas partners from the Danube region were not expected to take considerable part, let alone fully fledged collaborative projects. Nevertheless, we learnt that transport problems and their solutions played one of the major and integrative roles in the Danube Strategy, so we tried to relate that to our field of expertise – language technologies.

In January 2014 at the Networking Day on Horizon 2020 Work Programme 2014-2015 and Connecting Europe Facility in Luxembourg, the partners convened again and we got acquainted with the CEF programme there. Trying the CEF programme as one of the possibilities was initiated by Tamás Váradi, thus he suggested applying to its calls for project proposals beside the H2020 calls that we had been familiar with since we considered them as the prolongation of FP7. The CEF programme predominantly funded the development of systems that were regarded as

[1] Formally not a member of CESAR consortium, but he belongs to the Danube region.

DOI: https://doi.org/10.18135/VT70.15

enhancement to the connections between EEA countries such as transport infrastructure (e.g. intelligent transport systems (ITS), multimodal transport etc.). The CEF Telecom subprogramme, however, focused mainly on the telecommunication infrastructure (e.g. broadband internet connection), therefore a niche for LT opened within the framework of Digital Service Infrastructures (DSIs) where machine translation (MT) became one of the horizontal services that would ensure the multilingual usage of other DSIs.

Throughout our discussions about the possible research topics that would fit into the scope of the first calls and knowing that transport also played an important role not just in the Danube Strategy, but also in CEF, an idea to combine the ITS with LT crystallized in us in the form of a proposal for a service that would offer the multilingual approach to road traffic information. Such information is usually broadcast in the most natural and desired way – as spoken news about road conditions, preferably in the driver's native language.

2. ITS and multilingualism

The issue of language in real time traffic information is a challenge that seemed to be neglected in both the practice and at the policy level in ITS. It seemed obvious that regardless of the source of data and the technology of collecting and presenting them, seamless cross-border traffic information could not be delivered without overcoming language barriers. These impediments have been unable to be solved even within EU member states of Europe where administrative borders have disappeared. These multilingual requirements for a seamless cross-border service were not sufficiently addressed in strategic ITS documents.

Language-independent solutions such as pictograms have their place but their expressive power can't reach that of a natural language, let alone the native language of the driver. It can be stated with justification that ITS Deployment Guidelines seek to eliminate the problem of multilingualism rather than provide a solution (Váradi et al., 2015a).

3. Road Traffic Information Systems

Road traffic information systems still operate in a fragmented manner, isolated by language and even by country borders. Traffic-related information is provided by various local partners, while data collection and processing are carried out by national Traffic Information Centers (TIC).

Users can access information on the website of the national TIC, via mobile apps, through a call centre, or by listening to radio and TV broadcasts. However, the most commonly available channel is public radio broadcasting traffic information relevant to the whole country. Surveys indicated that this is indeed the most popular source and medium of traffic information, which drivers insist on using.

Most countries operate a road traffic information service on a national scale. Cooperation between them is limited. Seamless cross-border traffic information provision still remains only a long-term objective. The potential for data ex-change between the national services has been greatly facilitated by the development of a standardized notation and ontology in the transport domain, namely DATEX II. Nonetheless, no similar progress has been made to remove the linguistic barriers to streamline cross-border traffic information services. Real-time road traffic information services have typically been provided in the official language of a country, accompanied by occasional broadcast of limited scale and/or limited time in a foreign language, i.e. information for tourists during summer in English.

Road users crossing countries normally do not speak the language of the neighbouring country, thus they are isolated in a foreign language medium. If drivers do not understand traffic-related news abroad they are prone to get delayed in congestions, or they may even get involved in accidents in more serious cases (Váradi et al., 2015a).

4. MORENA as a possible solution

The solution to the problem of multilingual traffic information is not to avoid the production of natural language messages, but to deliver the information to the road users in languages they understand and in the most natural way they prefer, i.e. spoken native language. Language technology has reached such a level of maturity that it can offer a comprehensive full-scale solution to this challenge through a combination of real-time machine translation (MT) and natural-sounding speech technology.

We proposed the development of a robust, high quality MT system and its deployment in the ITS domain. We put our confidence in the unique infrastructure of the elaborate terminology and ontology (DATEX II) available for the ITS domain. This language-independent ontology represents a comprehensive and fine-grained conceptual system that allows the description of practically any traffic event and condition.

The task of machine translation should boil down to converting the real-time traffic information into a standard DATEX II representation, which can then be mapped into any particular target language and delivered through a text-to-speech system. The technology was expected to use proven components and promised to yield much higher quality translation than what was available at that time through freely accessible general purpose SMT methods (Váradi et al. 2015a).

Today, this effort would most probably include NMT methods, but it should be investigated in advance whether the DATEX II could function as controlled interlingua (glass box) or it should be left to neural networks themselves to form the connection network (black box).

The envisaged technology at that time presented a ground-breaking solution to pre- and on-trip traffic information provision by delivering traffic-related information in the most user-friendly manner, i.e. in the form of spoken messages in one's native tongue, something like an automated traffic news internet radio. It was planned to offer high quality MT through a hybrid approach integrating proven concepts in the fields of machine translation (MT), terminology management, computer-assisted translation using translation memories (TM), and controlled-natural language (CNL) systems. The LT was planned to draw on the standardized data dictionaries and protocols developed in the ITS domain and thus it would have utilized the synergies between the two disciplines involved (Váradi et al. 2015a).

In the first half of 2014, a detailed project proposal was developed to implement the technology described above. We called it MORENA. The backbone of the whole solution was a cloud-based system that would provide the machine translation of traffic information. This was called the MORENA service, which would allow the deployment in a variety of settings. As one possible implementation, the project offered to develop a multi-platform mobile application for mobile devices (smartphones, tablets, etc.), which we referred to as the MORENA application or MORENA app. The MORENA application was not intended to be the sole deployment of the MORENA service and the MORENA service should be assessed in terms of the potential it could offer as embedded technology in the ITS domain and should not be evaluated on the merits of the MORENA application alone (Váradi et al. 2015a). At that time we even had the idea to propose the MORENA service to become one of the CEF DSIs since it could have played an integrative role within the CEF Transport and/or CEF Telecom.

Then, the anticipated deployment of the suggested service was innovative as well. It was supposed to be a cloud-based service, delivered as an application for mobile devices. The drop in roaming charges regarding EEA countries by mid-2017 represented a major breakthrough, enabling mobile devices to become the main channel for broadcasting traffic information services (TIS). Such an online service would have presented the additional benefit over current practice because the information flow through this channel can be tailored to individual users since all relevant information such as preferred language, current GPS position, planned destination, etc. were at disposal from the device itself. Accordingly, delivery of information would have been personalized not only for language, but also for GPS position and destination (Váradi et al., 2015a).

5. Project presentations and proposals

Keeping an eye on the CEF programme, we still decided to submit the MORENA project proposal to the H2020-ICT-2014-1 call in April 2014. Unfortunately, this proposal was not successful, however, that didn't stop the consortium from further developing the idea and trying again.

We tried to look for additional opportunities, so Tamás Váradi and Marko Tadić also attended the Danube Region Transportation Days in October 2014 in Ljubljana, where a presentation was given on the MORENA project as an already established one, with its proposal submitted, as well as the first leaflet produced.

Figure 1. MORENA leaflet at the Danube Region Transportation Days, Ljubljana, 2014-10-21.

In early 2015 we submitted a more refined proposal again, but this time to the CEF Transport Multiannual Call: Specific Call for Cohesion Funds under the Transport Sector. The Funding Objective 2 in this call was defined as "Deployment of new technologies and innovation in all transport modes, with a focus on decarbonisation, safety and innovative technologies for the promotion of sustainability, operation, management, accessibility, multimodality and efficiency of the network". We expected the MORENA project to qualify for "new technologies and innovation" as well as for "safety and innovative technologies". The consortium was enlarged with respective national players in TIC who were included as providers of real-time data on national road traffic conditions. We also expected that their involvement as key stakeholders would help the project to be accepted. However, this proposal was evaluated by experts in transport and they didn't find that the deployment of LT in ITS would be worth-supporting at this stage.

We began to prepare for the CEF Telecom call in 2016, but the consortium was successful with another proposal, the MARCELL project, so the idea behind MORENA was gradually put aside.

Nevertheless, the work on this idea resulted in two papers presented in October 2015 at the 15th ITS Congress held in Bordeaux (see Varádi et al., 2015a and 2015b). Today, these two papers may be considered as pioneering papers towards introducing the usage of language technologies in ITS. In spite of all efforts, they have remained to represent the untapped potential behind the meticulous pioneering work spearheaded by Tamás Váradi.

6. Conclusions

On this solemn occasion we present the activities based on the idea and several proposals for the MORENA project in which our celebrated person was involved in the manner he has always pursued – entirely, thoroughly and with the best interest of the consortia.

Natalis honorem Tamasi Varadii septuagesimum.

References

EasyWay. Data Exchanges DATEX II Supporting Guideline (2012). Available at http://www.rits-net.eu/uploads/media/EW-DG-2012_DTX-DG01_DatexII_02-00-00.pdf

Freudenstein, J., Cornwel, I.: Tailoring a reference model for C-ITS architectures and using a DATEX II profile to communicate traffic signal information, Transport Research Arena 2014, Paris (2014)

Gilka, P., Richter, T.: Result assessment for user acceptance and safety evaluation on motorways with I2V-communication. In Proceedings of the 18th World Congress on ITS, Orlando, FL, USA (2011)

ITS Deployment Guidelines Library (2012). Available at https://dg.easyway-its.eu/DGs2012.

Váradi, T. Tadić, M. Gulyás, A. Niculescu, M.: Language Technology in the Service of Intelligent Transport Systems, Paper number ITS-2878, 22nd ITS World Congress, Bordeaux, France, 5–9 October 2015 (2015a)

Váradi, T., Tadić, M., Gulyás, A., Niculescu, M.: When Will ITS Speak Your Language? Paper number ITS-2955, 22nd ITS World Congress, Bordeaux, France, 5–9 October 2015 (2015b)

MARCELL – A project to remember: hard work of a friendly consortium under wise coordination

Dan Tufiș[1], Vasile Păiș[1], Verginica Barbu Mititelu[1], Radu Ion[1], Elena Irimia[1], Andrei Avram[1], Eric Curea[1]

[1] Institutul de Cercetări pentru Inteligență Artificială "Mihai Drăgănescu"
{tufis, vasile, vergi, radu, elena, eric}@racai.ro
avram.andreimarius@gmail.com

1. Collection and Annotation of the Romanian Legal Corpus

In this section we review the results of the Romanian team in the first part of the MARCELL project (https://marcell-project.eu/) whose ultimate goal was to enable the enhancement of automatic translation in CEF.AT[1] on the body of national legislation in seven EU official languages. For this task, all the seven teams from Bulgaria, Croatia, Hungary, Poland, Romania, Slovakia, and Slovenia cooperated in order to produce a comparable corpus heavily annotated (part-of-speech, syntactically parsed and semantically labelled by EUROVOC[2] identifiers and IATE recognized terms[3] appropriately marked-up).

EuroVoc is a multilingual thesaurus which was originally built up specifically for processing the documentary information of EU institutions. The covered fields encompass both European Union and national points of view, with a certain emphasis on parliamentary activities. The current release of EuroVoc (4.4) was published in December 2012. The new edition was the result of a thorough revision, among others, according to the concepts introduced by the Lisbon Treaty. It includes 6,883 unique IDs for thesaurus concepts (corresponding to the preferred terms), classified into 21 domains (top-level domains), further refined into 127 subdomains. Preferred terms and different variations of the preferred terms are assigned the same ID, subdomains and top-level domains.

IATE ('**I**nter**a**ctive **T**erminology for **E**urope') is the EU's terminology database. It has been used in the EU institutions and agencies since summer 2004 for the collection, dissemination and management of EU-spe-

[1] https://ec.europa.eu/inea/sites/inea/
[2] https://eur-lex.europa.eu/browse/eurovoc.html
[3] https://iate.europa.eu/home

DOI: https://doi.org/10.18135/VT70.16

cific terminology. It contains over 8 million terms in 24 official languages of the EU. The IATE database contains about 55,000 terms in Romanian.

The language-specific corpora were cross-lingually aligned at the top-level domains identified by EUROVOC descriptors. A general view of the project activities is given in another article (Váradi et al., 2020). As for the Romanian language, the current legal database includes more than 144k processed legislative documents. There are five main types of Romanian legal documents: governmental decisions (25%), ministerial orders (18%), decisions (16%), decrees (16%) and laws (6%). After the statistics were calculated, we found that there were six main issuers of the documents: Government (28%), Ministers (19%), President (14%), Constitutional Court (12%), Parliament (6%) and National Authorities (4%). Concerning the timestamp, most of the published documents were issued after 2000. Almost 4,000 documents were issued before 1990, and around 21,000 legal documents were published between 1990 and 2000. Following 2000, the number of issued documents increased. On average, more than 6,000 documents were issued every year, reaching a total of 120,000 until 2018, in 19 years. In terms of document length, there are around 6,000 short documents (less than 100 words per document, most of them being updates to other previously published legal documents), 70,000 documents contain between 100 and 500 words per document, more than 18,000 documents have around 1000 words per document and 52,000 contain more than 1000 words.

2. Linguistic Annotation

The corpus is annotated in batches as new documents are collected. All partners produced processing flows that were dockerized and stored as "ready-to-use" on the RELATE portal (Păis et al., 2019). The processing flows include language specific text normalization, sentence splitting, tokenization, POS tagging, lemmatization, dependency parsing, named entity recognition and classification, chunking, IATE term annotation and top level EUROVOC labeling.

The Romanian preprocessing pipeline, excluding IATE and EUROVOC annotations, is performed using the TEPROLIN text preprocessing platform (Ion, 2018). TEPROLIN offers the user various choices for each processing step and can be easily configured to different specific algorithms. It only needs a list of desired text annotations to infer and construct the pipeline getting these annotations out. TEPROLIN includes

mostly tools developed by our institute (e.g. TTL, MLPLA, NER, BIOner, Diac, TextNorm), however, not exclusively: we incorporated some other open-source algorithms (such as UDpipe, NLP-cube, Korap) into the preprocessing platform and will continue to add new better algorithms as they become freely available.

Dependency parsing is produced by NLP-Cube (Boros et al., 2018) which, according to the evaluations done in the CoNLL 2018 shared task "Multilingual Parsing from Raw Text to Universal Dependencies", has a labelled attachment score of around 85% for Romanian.

3. Automatic Identification of Legal Terms in Romanian Law Texts

As specified in the Grant Agreement of the project, each language-specific corpus was enriched with IATE and EUROVOC labels, then classified and multilingually clustered based on these annotations.

For term identification in both IATE and EuroVoc, the Romanian team used an algorithm similar to the Aho-Corasick algorithm (Aho and Corasick, 1975), using a language specific calibrated compressing function largely described in (Coman et al., 2019). This method only implies linear-time transformations of the IATE dictionary (through the compression function) and a single pass through the Aho-Corasick structure, the overall complexity of the proposed algorithm is linear and has a term matching rate of approximately 98%. Besides the compression function, the Aho-Corasick structures were not language-specific and were created during runtime (for the Romanian IATE terms, consisting of about 55,000 terms, the computation time was approximately 10 seconds).

Thus, the algorithm would be available for any language, provided that a specific compression function relevant to that respective language was accessible.

The processing allowed the annotation of the corpus through the introduction of EuroVoc and IATE labels. Thus, every occurrence of the IATE term inside the corpus is now annotated by its respective position and is accompanied by the corresponding EuroVoc categories. In the analysed testing sample, we were able to detect no false positive matches and a nearly perfect precision for detecting true matches. The overall matching rate over the used testing samples was approximately 98%-99%, however, as mentioned earlier in the paper, we expect the real matching rate to have a slightly lower value due to unexpected collisions

which may have occurred during the term identification process. Moreover, the matching rate, as defined in the paper, has a rather simple definition and it presents the accuracy of our work only to some extent.

Secondly, by using the aforementioned annotation with EuroVoc categories, we were able to create a statistical database documenting the occurrence frequency of all the categories in each legal document. For each file, we determined the number of terms falling in each of the 21 EuroVoc categories. After multiplying each frequency with a predefined weight, the mentioned file was placed in the category corresponding to the maximum number of terms.

The predefined weights were roughly determined over a testing sample of medium size (approx. 100 documents), due to the lack of pre-processed legal data. As most of the EuroVoc categories present in the description of the IATE terms yielded correct classifications, we only needed to slightly modify some of them (such as Geography) to obtain a better classification over the testing sample. This simple classification method was replaced later with a more sophisticated one taking advantage of available word embeddings (see further).

The computation time for simply identifying the matches with the hybrid algorithm was approximately 4,250 seconds, which yields a rate of almost 35 documents per second. This experiment was performed using a server with two Xeon 4210 CPUs at 2.2 GHz with 20 annotation threads, yielding over 1 document per second considering a single annotation thread. Thus, as the XML-formatted corpus had a size of approximately 31.2 Gb, the processing rate was 7.5 Mb of text per second. Because of using the Aho-Corasick structure, the memory usage was also linear, which made the computation possible on almost any machine. The number of matches was significantly increased by working with the lemmatised corpus instead of the unlemmatised one. By using the described algorithm, we have identified a total of 51,517,877 matches (IATE terms), out of which 29,162,667 were short terms (single word terms) and 22,355,210 were long terms (multiple word terms). The term identification step is based on the encoded list of IATE terms, which brings our approach closer to a gazetteer-based processing. This is why the estimated precision is so high.

4. Document classification and evaluation

In order to obtain an objective evaluation of our document classification algorithm it was necessary to have reference data, classified and validated by human experts. These requirements were not met for the MARCELL Romanian corpus, but instead we resorted to a well known multilingual corpus, JRC-Acquis (Steinberger et al., 2006), which was processed by a publicly available program called JEX.

JEX (Steinberger et al., 2012) is a multi-label classification software developed by JRC, trained to assign EuroVoc descriptors to documents. Its primary concern was to cover the activities of the EU. Written using Java, it provides scripts for pre-processing a collection of documents, training a new model, post-processing the results and evaluating a new model. Each script employs a configuration file for the required parameters. The toolkit also comes with a graphical interface (GUI) for users to label new text, XML, HTML documents or to interface with training scripts for obtaining a classifier on their own documents. However, the usage of the GUI interface is optional and the toolkit allows for simple command line execution over collections of documents.

Based on (Pouliquen et al., 2003), JEX classification algorithm relies on a list of lemma frequencies obtained from normalized text together with associated weights, statistically related to each descriptor. These are called associates or topic signatures. At runtime, given the new document's list of lemma frequencies the algorithm picks the descriptors of the associates that are the most similar to it. The JEX package offers pre-trained classifiers for 22 official EU languages, including Romanian (trained on over 25,000 documents, consisting of manually annotated ACQUIS and OPOCE corpora). (Steinberger et al., 2012) reports a F1 score of 47.84% (derived from P=45.55% and R=50.43%, computed for predicting 6 EuroVoc identifiers).

More recently, researchers tried to further improve the perfomance of JEX regarding different languages. For instance, the Italian language, Boella et al. (2012), mono-label transformations (Tsoumakas and Katakis, 2007), and employing Support Vector Machines (SVM) (Joachims, 1998) for classification, achieves an F1 score of 58.32%. While applying the Croatian CroVoc (an extended EuroVoc terminology) on the NN13205[4] corpus (different from the JEX training and testing sets), Šarić et al. (2014)

[4] http://takelab.fer.hr/data/nn13205

reports an F1=68.60%. We are unaware of other studies regarding EuroVoc classification for Romanian, therefore JEX is the only available tool/algorithm.

5. Dataset used for the evaluation exercise, the word-embeddings and the method

After downloading and extracting the JEX extended package,[5] we are presented with the corpora on which it was trained, consisting of the ACQUIS and OPOCE corpora. JEX used a regular cross-validation approach for evaluation, consisting of creating multiple splits out of the training data and evaluating each one followed by averaging the results. However, the individual splits are not provided, therefore we had to create our own splits. Before doing this, we first annotated the corpora using our RELATE platform with a pipeline similar to the one used for the Marcell project. Furthermore, the platform was applied to extract statistics on the corpora. Finally, a script took care of creating 10 split folders, each consisting of 80% training data and 20% test data with gold annotations from the original corpora. Furthermore, the 80-20 split rule was applied on both corpora, thus producing balanced splits.

Word representations learned using artificial neural network approaches (Mikolov et al., 2013) have previously been used successfully in a number of natural language processing tasks, including classification (Joulin et al., 2017). However, this had not been applied to EuroVoc classification. Facebook Research introduced the FastText[6] tool initially intended for training neural embeddings together with sub-word information (Bojanowski et al., 2017). Using this tool, we previously created and evaluated word representations on the Reference corpus for Romanian language CoRoLa (Barbu Mititelu et al., 2018). These results were reported by Păiș and Tufiș (2018) and can be freely downloaded from the website of our Institute.[7] An advantage of word embeddings representation is that once trained and evaluated, these representations can be used directly for converting words into numeric (floating point) vectors, suitable as input to other algorithms. This ensures a starting point given by the accuracy of the word representation and reduces the time needed for training more advanced algorithms.

[5] https://ec.europa.eu/jrc/en/language-technologies/jrc-eurovoc-indexer#Download%20JEX
[6] https://fasttext.cc/
[7] http://corolaws.racai.ro/word_embeddings/

The FastText tool was further enhanced, enabling training a linear classifier based on word embeddings and encoding of input documents. Therefore it seemed like an obvious choice for using the previously generated representations to try and classify texts using the EuroVoc terminology. The tool allows for adapting the model parameters to a specific language by considering the minimum and maximum lengths of character and word n-grams. Additionally, other parameters such as learning rate can be further fine-tuned.

For each of the previously created splits, a Romanian language classifier was trained. Then it was evaluated on each of the test corpora and the results were finally averaged to produce the final data (similar to the JEX evaluation approach). This allowed us to obtain an average F1=53.53% (compared to the JEX reported F1 of 47.84%, this gives us an increase of 5.7%). Similarly, we noted increased performance for both precision (50.93% our result compared to 45.17% from JEX) and recall (56.41% our result compared to 50.19 from JEX).

For the purposes of the Marcell project, we further converted the EuroVoc identifiers into MT labels and finally top-level domains. This is possible given that the mapping is present in the EuroVoc. There is a direct mapping from an identifier to an MT label and further to a top-level domain, represented by the first two digits of the MT label. Reverse mapping is not possible directly, since multiple identifiers are associated to a MT label.

In the context of the Marcell project, documents are classified using only EuroVoc top-level domains. In order to give an estimate of our classifier at this level we converted both the gold corpora annotations and the classifier automatic annotations to top-level annotations (considering both our approach and the JEX approach). We then applied again the scoring algorithm on all the splits and computed a final average over all the splits. This produced a top-level F1 score of 70.80% for our approach, compared to a score of 64.88% for JEX, thus providing an increase of almost 6% (comparable, yet slightly better than the identifier based evaluation). Similarly, we noted increases in both precision (64.90% our method vs 59.34% JEX) and recall (77.89% our method vs 71.56% JEX).

After performing the evaluations, a classifier was trained on the entire training corpora, providing a model which should have similar performances to the reported averaged ones (however in this case no further evaluation can be performed since there is no additional data to compare against). This final model was used to classify the Romanian Marcell

legislative corpus. This produced a rather unbalanced distribution of top-level domains with more than 80,000 documents assigned to the domains geography (72) and European Union (10). At the other end, less than 1,000 documents were assigned to the domains international organisations (76), science (36) and industry (68).

Apart from the EuroVoc classification, the Marcell corpora annotations include term identification. There are three options for this purpose: identifiers, MicroThesaurus (MT) labels or top-level domains. We consider MTs to be more useful for tasks like multilingual clustering, which was one of the goals of the project. This happens because MT labels include semantic information as opposed to the identifiers, which are used only as record ID in terminology. Furthermore, the large number of identifiers (6883), compared to 127 MT labels makes it more challenging for cross-lingual clusterization to decide identifier similarity (possibly requiring additional processing), while the MTs already utilize the hierarchical nature of the EuroVoc terminology. Finally, the MT architecture is more stable to changes in terminology, contrary to the identifiers which are growing in number or get removed (as certain terms may become obsolete).

The new EuroVoc classification method for Romanian legal documents presented in this section was integrated in the RELATE platform. First, it is possible to annotate single documents.[8] In this mode, the user can enter the text document, the number of identifiers to be predicted and a threshold for the identifier association probability. By default, the number of predicted identifiers is set to 6 and the threshold to 0. This corresponds to the same values used during the JEX comparison. The platform page is presented in the following image.

[8] https://relate.racai.ro/index.php?path=eurovoc/classify

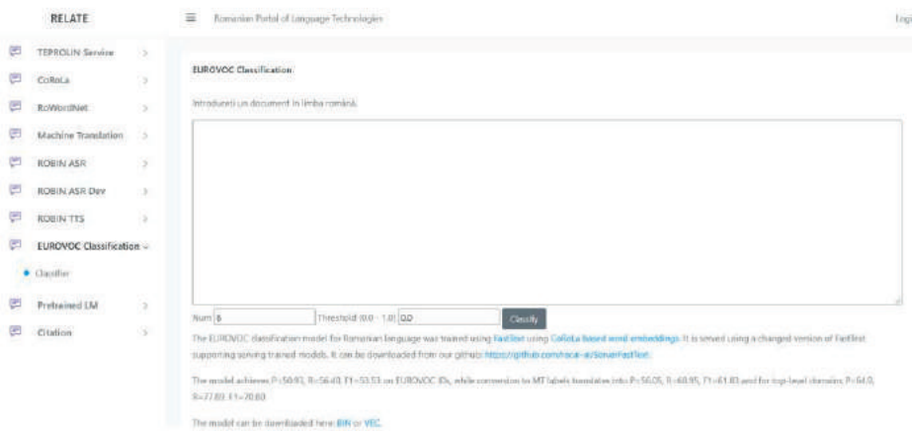


Figure 1. Single document EuroVoc classification in the RELATE platform.

Following the execution through the system, the platform presents the associated identifiers. These are then converted into MT labels and finally into top-level domains. Depending on the document entered, the number of identifiers is usually larger than the number of MT labels, which in turn is larger than the number of top-level domains. An example is presented in the next image. The transformation is handled automatically within the platform, using the EuroVoc hierarchy.

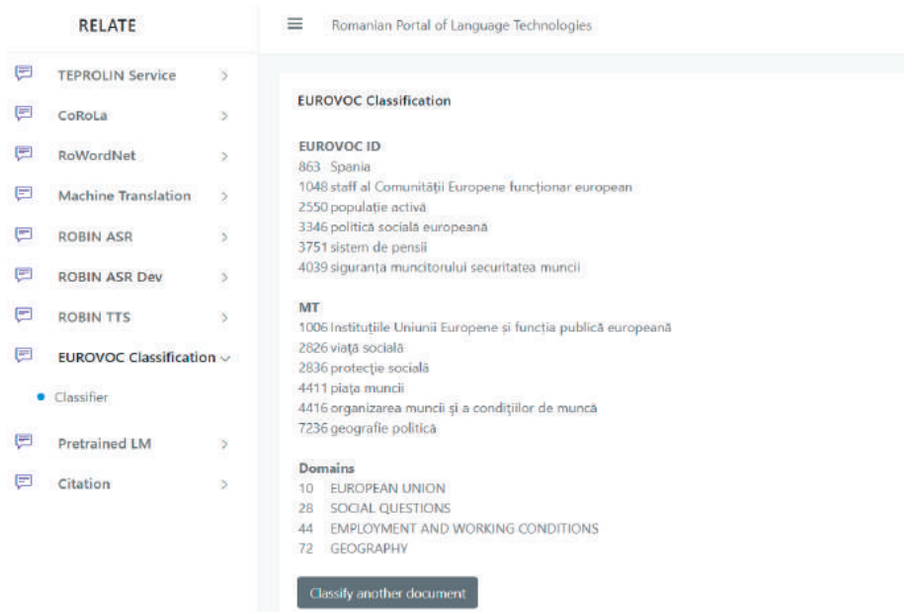


Figure 2. Results from single document EuroVoc classification in the RELATE platform.

The second integration is realized in the internal part of the platform, used for corpora annotation. This already provides mechanisms for uploading large corpora and performing sentence splitting, tokenization, part-of-speech tagging, dependency parsing. Furthermore, the integration of EuroVoc classification is available at the end of the processing pipeline as an additional task. Invocation of the new task is presented in Figure 4.

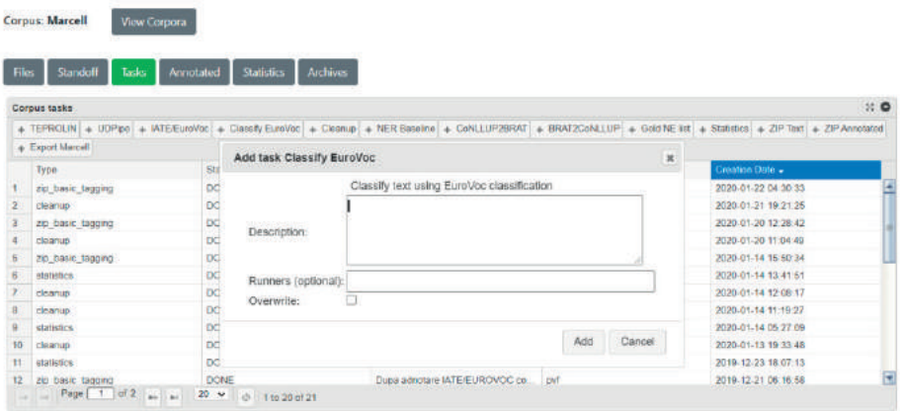


Figure 3. Corpora processing integration in the RELATE platform.

According to the Marcell specification, the results of EuroVoc classification is available in the metadata fields of each annotated file. Since we use CoNLLU-Plus format for the annotations, we first defined the columns like this "# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC MARCELL:IATE MARCELL:EUROVOC", thus considering the last two columns to correspond to IATE and EuroVoc terms. The EuroVoc classification is given by the line "# eurovoc_domains = 04 08 10 24". An example is presented in the following figure.

```
# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC MARCELL:IATE MARCELL:EUROVOC
# eurovoc_domains = 04 08 10 24
# sent_id = 1
# text = ORDIN nr. 3.154 din 24 octombrie 2008 pentru modificarea si completarea Normelor metodologice de aplicare a prevederilor Ordonantei
1   ORDIN   ordin   ADP   Spsa   AdpType=Prep|Case=Acc   2   case   _   SpacesBefore=\n   _   _
2   nr.   nr.   NOUN   Yn   Abbr=Yes   0   root   _   _   _   _
3   3.154   3.154   NUM   Mc-p-d   Number=Plur|NumForm=Digit|NumType=Card   2   nummod   _   _   _   _
4   din   din   ADP   Spsa   AdpType=Prep|Case=Acc   6   case   _   _   _   _
5   24   24   NUM   Mc-p-d   Number=Plur|NumForm=Digit|NumType=Card   6   nummod   _   _   _   _
6   octombrie   octombrie   NOUN   Ncms-n   Definite=Ind|Gender=Masc|Number=Sing   2   nmod   _   _   _   _
7   2008   2008   NUM   Mc-p-d   Number=Plur|NumForm=Digit|NumType=Card   6   nummod   _   SpacesAfter=\n   _   _
8   pentru   pentru   ADP   Spsa   AdpType=Prep|Case=Acc   9   case   _   _   _   _
9   modificarea   modificare   NOUN   Ncfsry   Case=Acc,Nom|Definite=Def|Gender=Fem|Number=Sing   2   nmod   _   _
10   si   si   CCONJ   Crssp   Polarity=Pos   11   cc   _   _   _   _
11   completarea   completare   NOUN   Ncfsry   Case=Acc,Nom|Definite=Def|Gender=Fem|Number=Sing   9   conj   _   _
12   Normelor   normă   NOUN   Ncfpoy   Case=Dat,Gen|Definite=Def|Gender=Fem|Number=Plur   11   nmod   _   _   _
13   metodologice   metodologic   ADJ   Afpfp-n   Definite=Ind|Degree=Pos|Gender=Fem|Number=Plur   12   amod   _   _   _
14   de   de   ADP   Spsa   AdpType=Prep|Case=Acc   15   case   _   _   _   _
```

Figure 4. A Romanian legal document annotated according to Marcell specifications.

6. Marcell project sustainability

Most projects end providing new data once their allocated project duration expires. In the case of Marcell we considered a sustainability scenario in which new data could be provided even after the project ended. In order to ensure future operation, each annotation pipeline was embedded into a Docker container with all the required resources. Furthermore, a basic graphical user interface (GUI) was constructed in the form of a web site. This GUI was also dockerized with all the configuration and needed resources. The entry point of the Marcell sustainability GUI is presented in Figure 5.

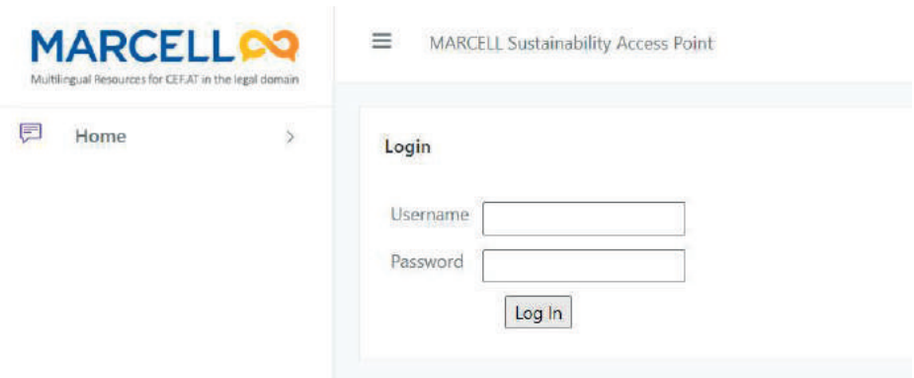


Figure 5. Marcell sustainability GUI entry point.

In order to construct the sustainability framework, each partner provided docker scripts for their pipelines. These were initially staged on the RELATE server and tested. The GUI itself was initially a stripped down version of RELATE which was further augmented with Marcell specific customizations as well as integration of the different language specific pipelines. The resulting platform allows uploading new raw text archives (as new legislation becomes available), then starting an annotation process which calls the corresponding pipeline depending on the specified language. Depending on the number of dockers available for each language, the pipeline invocation can happen in parallel, thus reducing the overall time required for annotation. Finally, the results are stored back in the GUI and can be exported in the MARCELL specific format. Figure 6 presents different corpora loaded in the GUI during the sustainability framework testing.

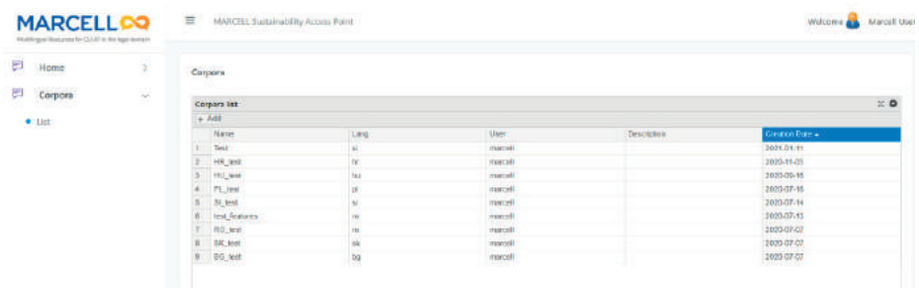


Figure 6. Corpora loaded in the Marcell sustainability framework GUI.

The GUI offers additional functionality such as computing statistics, allowing to see how the corpus grows over time, archiving of raw and annotated text, allowing to store different versions of the corpus. Resulting archives can be downloaded as ZIP files containing Marcell formatted documents.

6. Instead of Conclusions

The MARCELL consortium includes some veterans of the CEE Language Technology area, who have known and respected each other for more than 20 or 30 years. They are accompanied by younger and experienced researchers who will take over the responsibilities for the technologies of our national languages and accomplish the roadmap for all European Language Equality. Dr. Tamás Váradi has led the way for many years in a professional and elegant manner, being attentive to all details implied by his coordination of large scientific consortia. Dr. Tamás Váradi is an ideal project coordinator, diligent to observe the deadlines and milestones, harmonize efforts of his partners towards a successful accomplishment of the objectives.

The ICIA team wishes him a long and prosperous life with many more accomplishments.

A big and whole-hearted THANK YOU, Tamás!

References

Aho, A., Corasick, M.: Efficient string matching: An aid to bibliographic search. Commun. ACM 18/6, 333–340 (1975)

Barbu Mititelu, V., Tufiș, D., Irimia, E.: The Reference Corpus of the Contemporary Romanian Language (CoRoLa). In: Calzolari, N., Choukri, Kh., Cieri, Ch., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (eds.) Proceedings of the International Conference on Language Resources and Evaluation. pp. 1178–1185 ELRA (2018)

Boella, G., Di Caro, L., Lesmo, L., Rispoli, D.: Multi-label Classification of Legislative Text into EuroVoc. In: Schäfer, B. (ed.) Legal Knowledge and Information Systems: JURIX 2012: the Twenty-fifth Annual Conference Vol. 250. pp. 21. IOS Press, Amsterdam (2012)

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching Word Vectors with Subword Information. In: Transactions of the Association for Computational Linguistics Vol. 5. pp. 135–146 (2017)

Boroș, T., Dumitrescu, S. D, Burtică, R.: NLP-Cube: End-to-End Raw Text Processing With Neural Networks. In: Zeman, D., Hajič, J. (eds.) Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. pp. 171–179. Association for Computational Linguistics, Brussels, Belgium (2018)

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P.: Natural language processing (almost) from scratch. Journal of machine learning research 12, 2493–2537 (2011)

Coman, A., Mitrofan M, Tufiș D.: Automatic identification and classification of legal terms in Romanian law texts. In: Onofrei, M., Bibiri, A-D., Dragoș Nicolae, C., Tufiș, D., Cristea, D. (eds.) Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language (ConsILR 2019). pp. 39–49. Editura Universității "Alexandru Ioan Cuza", Iași (2019)

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning Word Vectors for 157 Languages. In: Calzolari, N., Choukri, Kh., Cieri, Ch., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (eds.) Proceedings of the International Conference on Language Resources and Evaluation (2018)

Ion, R.: TEPROLIN: An Extensible, Online Text Preprocessing Platform for Romanian. In: Păiș, V., Gîfu, D., Trandabăț, D., Cristea, D., Tufiș, D. (eds.) Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language (ConsILR 2018) Editura Universității "Alexandru Ioan Cuza", Iași (2018)

Ion. R.: Word Sense Disambiguation Methods Applied to English and Romanian. PhD Thesis. pp. 148. Romanian Academy, Bucharest (2007)

Joachims, T.: Text categorization with Support Vector Machines: Learning with many relevant features. In: Machine Learning: ECML-98. pp. 137–142. Springer Verlag, Berlin, Heidelberg (1998)

Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of Tricks for Efficient Text Classification. In: Lapata, M., Blunsom, Ph., Koller, A. (eds.) Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics Volume 2, Short Papers. pp. 427–431. Association for Computational Linguistics, Valencia, Spain (2017)

T. Mikolov, K. Chen, G. Corrado, J. Dean: Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 (2013)

Paiș, V., Tufiș, D.: Computing distributed representations of words using the CoRoLa corpus. In: Proceedings of the Romanian Academy. Series A Vol. 19, No. 2, pp. 403–409. Romanian Academy, Publishing House of the Romanian Academy, Bucharest (2018)

Păiș, V., Tufiș, D., Ion, R.: Integration of Romanian NLP tools into the RELATE platform. In: Onofrei, M., Bibiri, A-D., Dragoș Nicolae, C., Tufiș, D., Cristea, D. (eds.) Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language (ConsILR 2019). pp. 181–192. Editura Universității "Alexandru Ioan Cuza", Iași (2019)

Pouliquen, B., Steinberger, R. and Ignat, C.: Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus. In: Cristea, D., Ide, N., Tufiș, D. (eds.) Proceedings of the Workshop Ontologies and Information Extraction at the Summer School The Semantic Web and Language Technology –Its Potential and Practicalities (EUROLAN 2003). Bucharest, Romania (2003)

Šarić, F., Dalbelo Bašić, B., Moens, M. F., Šnajder, J.: Multi-label classification of croatian legal documents using EuroVoc thesaurus. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of SPLeT-Semantic processing of legal texts: Legal resources and access to law workshop. ELRA, Reykjavik (2014)

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D. and Varga, D.: The JRC-Acquis: A multilingual aligned parallel corpus with 20+languages. In: Calzolari, N., Choukri, Kh., Gangemi, A., Maegaard, B., Mariani, J., Odijk, J., Tapias, D. (eds.) Proceedings of the 5th international conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy (2006)

Steinberger, R., Ebrahim, M. and Turchi, M.: JRC EuroVoc Indexer JEX - A freely available multi-label categorisation tool. In: Calzolari, N., Choukri, Kh., Declerck, T., Ugur Dogan, M., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) Proceedings of the 8th international conference on Language Resources and Evaluation (LREC 2012). pp. 798–805. Istanbul, Turkey (2012)

Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM) 3(3), 1–13 (2007)

Tufiș, D., Mitrofan, M., Păiș, V., Ion, R., Coman, A.: Collection and Annotation of the Romanian Legal Corpus. In: Calzolari, N., Béchet, F., Blache, P., Choukri, Kh., Cieri, Ch., Declerck, Th., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the 12th Language Resources and Evaluation Conference. pp. 2766–2770. European Language Resources Association, Marseille, France (2020)

Tufiș, D., Mititelu, V. B., Irimia, E., Păiș, V., Ion, R., Diewald, N., Mitrofan, M., Onofrei, M.: Little strokes fell great oaks. creating CoRoLa, the reference corpus of contemporary Romanian. In: Revue roumaine de linguistique no.3. pp. 227–240. Editura Academiei Romane, Bucarest (2019)

Váradi, T., Koeva, S., Yamalov, M., Tadić, M., Sass, B., Nitoń, B., Ogrodniczuk, M., Pęzik, P., Verginica, B.M., Ion, R., Irimia, E., Mitrofan, M., Păiș, V., Tufiș, D., Garabík, R., Krek, S., Repar, A., Rihtar, M., Janez, B.: The MARCELL Legislative Corpus. In: Calzolari, N., Béchet, F., Blache, P., Choukri, Kh., Cieri, Ch., Declerck, Th., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020). pp. 3754–3761. Marseille, France (2020)

A gépi fordítás hetvenéves története

Prószéky Gábor[1]

[1] Nyelvtudományi Intézet
proszeky.gabor@nytud.hu

1. A hetven az hetven

Váradi Tamás hetvenedik születésnapja alkalmából arra teszek kísérletet, hogy a számítógépes nyelvészeti kutatások talán legismertebb, Tamással gyakorlatilag egyidős és az ő tevékenységei között többször is érintett területnek, a gépi fordításnak a hetvenéves történetét röviden összefoglaljam. Három nagy szakaszra szokás felosztani ezt az időszakot: a szabályalapú fordítás, a statisztikai fordítás és a neurális hálókkal történő gépi fordítás időszakára. Az első szakasz volt a leghosszabb, amely gyakorlatilag az ötvenes évek elejétől a géppel elérhető nagyméretű szövegkorpuszok megjelenéséig, a kilencvenes évekig tartott. Az ekkor megjelenő statisztikai közelítések egészen a neurális hálós módszerek megjelenéséig, a most befejeződött évtized elejéig tartottak. Napjainkban az uralkodó tudományos paradigma két fontos ismérve, hogy nem a nyelvészeti, de sokszor még nem is a programozási tudás az, ami a fordítási minőséget jelentős mértékben képes feljavítani, hanem a neurális rendszerek egyfajta „paraméter-beállítási" intuíciója. A tanítóanyagok gondos kiválasztásának és előfeldolgozásának megnőtt a jelentősége, ami ugyan igényli a tapasztalt nyelvész közreműködését, ám az új gépi környezetben sok, korábban jelentős eredményt elérő nyelvész a teljesen új szemlélet miatt kevésbé sikeres. Tamás azonban ebben a – hagyományos nyelvészeti orientációjú közelítéseknél sokkal több technikai ismeretet igénylő – világban, a neurális hálók világában is ígéretes első eredményeket tudhat magáénak.

2. A gép elkezd fordítani, a nyelvész szabályai alapján

A szélesebb értelemben vett *számítógépes nyelvészet* a számítógép és a nyelvészet számos lehetséges találkozási pontján kialakult szakterület. Ezen belül a *nyelvtechnológia* – azaz a mai tudományos világban hasz-

DOI: https://doi.org/10.18135/VT70.17

nált angol elnevezéssel: *human language technologies* – úgy definiálható, hogy ez az informatikának az az ága, ahol a nyelvészeti kutatásokon alapuló eredmények beépülnek a számítógépes rendszerekbe. Teszik ezt úgy, hogy a felhasználók számára a számítógéppel való kommunikáció folyamán az így kialakított szoftverrendszerek – bizonyos célhelyzetekben – a nyelvet jól használó emberéhez hasonló támogatást tudnak adni. Világosan kell látni, hogy eddig a nyelvet kizárólag az ember számára írta le a nyelvész, így bizonyos pontokon módja volt „összekacsintani" leendő olvasójával, építve arra, hogy az is ember, méghozzá nagy eséllyel hasonló kulturális háttérrel, így bizonyos alapvető fogalmak megmagyarázására nem volt szükség. A számítógép, amelynek számára leírjuk a nyelvet, nem rendelkezik azokkal a háttérismeretekkel, amivel egy nyelvtant értelmező ember, így minden olyan fogalmat, amelyre szükség lehet a rendszer működtetéséhez, a gépek számára részleteiben le kell írni. Egy egyszerű analógiával megvilágítva, ha a „vásárlást" mint tevékenységet leírnánk a gép számára, akkor azt a tényt, hogy a végén „oda kell menni a pénztárhoz", a gép csak úgy tudja értelmezni, ha az ehhez szükséges „menést" mint tevékenységet ismeri, különben kénytelenek vagyunk részletesen ezt is kibontani, azaz a „lépéseket" mint a „menés" alapelemeit is definiálni kell számára, és így tovább.

Napjainkra megjelent tehát egy új eszköz, mely az emberen kívül először képes a nyelvi leírás működtetésére: ez természetesen a számítógép, ami új nyelvészeti közelítések kialakítását is magával hozta. Talán a fentiekből az is világossá vált, hogy a 20. század közepétől kezdve ki kellett hogy alakuljon egy olyan nyelvleírási mód, amely csak részben azonos a nyelvészet addig meghatározónak tűnő elméleteivel, és sok olyan elemet tartalmaz, melyet az ember számára annak idején nem kellett leírni. A nyelvekkel kapcsolatban az általános tapasztalat ugyanis a 19. század közepéig az volt, hogy a nyelv változik. Ezért valójában a nyelvészet története a 20. századig elsősorban a történeti nyelvészet története volt. A 20. században megjelenő leíró, vagy más néven *deskriptív nyelvészet* viszont egyfajta „mechanikus" leírásnak is tekinthető, melyet már a számítógép létrejötte előtt egyfajta algoritmikus szemlélet jellemzett. A számítógépről ismeretes, hogy bizonyos értelemben a második világháború „hozadéka". Az eszköz neve igen sok nyelven a számolással, azaz a *comput-* latin tő valamely származékával kapcsolatos szóból alakult ki. Az egyik talán kevésbé ismeretes fő ok a számítógép létrejöttében a háborúkban oly fontos titkosírások mechanikus, sőt elektromechanikus kezelésének vágya, azaz a kódolás-dekódolás folyamatának gépesítése volt.

A világháború végén, a hidegháború kialakulásának hajnalán az Atlanti-óceán mindkét partján megjelent a gondolat, hogy a kódolás és dekódolás viszonya és az emberi nyelvek fordítása hasonló jellegű tevékenység, így egy ilyen eszköz létrejötte a gépi fordítás megvalósíthatóságának gondolatát is egyre erősítette (Hutchins, 1997). Ehhez nagy lökést adott, az MIT meghatározó hatású, kiváló matematikusának, Bar-Hillelnek az ötvenes évek elején tett kijelentése, miszerint idő kérdése csak, de a teljesen automatikus gépi fordítás megvalósítható (Bar-Hillel, 1951). Az Egyesült Államok kormánya nem kevés pénzt koncentrált erre az ígéretes kutatási területre, ami elsősorban az orosz műszaki-katonai szövegek fordításának automatizálását célozta meg. Az első működő gépi fordítást végző számítógép 1954-ben mutatkozott be az IBM georgetowni központjában (IBM, 1954). A fordításban részt vevő nyelvek leírása a gép számára azonban nem a nyelvészek által követett úton történt. Ennek egyik oka, hogy a gépi fordítást végző kutatók igazán nem is a nyelvet akarták leírni, hanem azt a módszert szerették volna megragadni, melynek segítségével az egyik nyelv szerkezeteit a másik nyelv szerkezeteivé tudja alakítani az ember. Az alapgondolat az volt, hogy ha ez a módszer megvan, akkor akár egy program is végre tudja hajtani a lépéseit. A fordítási egység a mondat volt, de nem abban a generatív értelemben, amelyről ebben az évtizedben már Chomsky egyre többet publikált (Chomsky, 1957). Ennek az egyik nyilvánvaló oka, hogy a Chomsky-modell az ideális beszélő nyelvi kompetenciáját volt hivatva megfogalmazni, a gépi fordításhoz pedig a bemenő mondatot egy nem feltétlenül ideális beszélő hozta létre, és a feldolgozás eredményeként sem egy absztrakt nyelvi szerkezetet, hanem egy másik nyelvi fordítást kellett a gépnek produkálnia. A világ akkori másik pólusán, a szovjet blokkban is folytak természetesen a kutatások, de a számítógépesítés alacsonyabb foka miatt a születendő nyelvi modelleket inkább matematikai nyelvészetinek nevezték (Papp, 1964). A Szovjetunió néhány neves nyelvészének hatására ezt követően az ún. szocialista országokban, így hazánkban is megindult a gépi fordítás kutatása. A gépi nyelvészet akkori amerikai eredményei – és nemcsak a „hivatalosan” publikáltak – ma is fellelhetők az Országos Műszaki Könyvtár által az ötvenes évek végén és a hatvanas évek elején beszerzett és félig-meddig titkos mikrofilmeken. A módszerek eleinte ugyan próbálták ötvözni az akkortájt születő generatív nyelvelméletek eredményeit a gépi feldolgozással, de egyre jellemzőbbé váltak nálunk is és máshol is a nyelvelméletmentes gépi kísérletek. Matematikai szempontból az volt az egyik probléma, hogy a Chomsky-féle

transzformációk nem invertálhatók. Ez számítógépes szempontból azt jelenti, hogy egy mondatátalakításkor kitörölt vagy elmozgatott elem helyét, az ún. nyomot a mondatelemző program nem találja meg, ui. a generatív levezetés végén ezek törlődnek. Az ilyen, a mondatban elvileg ott levő, de fizikailag nem megtalálható elemek visszaállítása az esetek jelentős részében nem, vagy csak nagyon hosszú idő alatt történhet meg. Már pedig komoly időbeli eltérés a mondatelemzés és a mondatlétrehozás között az emberi nyelvfeldolgozás esetén nem ismert, így furcsa volna egy olyan modell, mely egész máshogy működik generáláskor, mint elemzéskor. A számítógépes szakembereknek ugyanis elsősorban az emberek által létrehozott, és olykor nem pontosan megfogalmazott mondatokat kell elemezniük, és nem ideális mondatokat létrehozni. Így a számítógépes gyakorlatban egymás után jelentek meg olyan nyelvelméleti modellek, melyek nem a Chomsky-féle irányt követték, hanem például az őt megelőző strukturális leírást (Harris, 1951) vagy azt az alternatív elméletet, mely elsődlegesen a szavak közötti függőségi viszonyt szándékozott leírni (Tesnière, 1959). A teljesen automatikus gépi fordítás megvalósíthatóságát épp azok kezdték megkérdőjelezni az évtized végére, akik az évtized elején még az ügy élharcosai voltak, így a fordítással foglalkozó számítógépes kutatók elkezdtek a nyelv más, nemcsak fordítással kapcsolatos gépi feldolgozásával foglalkozni. Ekkor alakult ki az immár nemcsak a gépi fordítást magába foglaló számítógépes nyelvészet fogalma. Ehhez az Egyesült Államokban a gazdasági-politikai háttér is adott volt: a hidegháború eddig is a kutatási támogatás fő motiválója volt, de most már nemcsak a gépi fordításra koncentráltak. A Holdra szálláshoz például elkészült egy olyan számítógépes nyelvészeti program, amely a lehozott holdkőzetek adatbázisához angol nyelvű mondatok segítségével való hozzáférést biztosított (Woods, 1973). Ez bizonyos értelemben fordítóprogramnak volt tekinthető, bár a rendszer célnyelve nem emberi nyelv, hanem egy adatbázis-kezelő program nyelve volt. A számítógépes nyelvészet szempontjából lényeges, hogy Woods ennek a rendszernek a működtetéséhez létrehozta az Augmented Transition Network nevű leíró-működtető formalizmust (Woods, 1970). Az eljárás a véges állapotú automatáknak az emberi nyelvek rekurzív szerkezeteinek kezelésére is alkalmas kiterjesztésén alapult, és az ezt követő években, sőt évtizedekben a pszicholingvisztika és az amerikai számítógépes nyelvészet egyik alapmodelljévé vált, jóllehet visszalépéses

elven történő működése elsősorban a – gépi fordító rendszerek egyik állandó nyelve – az angol mondatainak feldolgozásakor volt csak evidens, a más típusú, például szabad szórendű nyelvek mondatainál nem.

A mesterségesintelligencia-kutatásból ekkortájt kinövőfélben levő nyelvvel kapcsolatos gépi alkalmazások másik legismertebbje Winograd nevéhez fűződik (Winograd, 1972). Ő a nyelv procedurális közelítésével kísérletezett. SHRDLU nevű rendszere egy olyan világot mozgat meg (angol) nyelvi instrukciók segítségével, melyben egy síklapon elrendezve háromdimenziós geometriai objektumok vannak csak, színükkel, méretükkel és alakjukkal. A nyelvi bemenet hatására a világ változásait reprezentálják, így az fizikai átrendezés ebben a virtuális világban a begépelt parancsok hatására megy végbe, amiről a gép „tud", és megfelelően reagál. Itt tehát a nyelv gépi reprezentációja procedurális, hiszen a nyelvi megnyilvánulások számítógép által végrehajtható műveletekbe való gépi fordításáról van szó.

Az eredeti értelemben vett gépi fordítás nagy túlélői viszont annak ellenére működtek, hogy az Egyesült Államok kormánya által a gépi fordítási eredmények – illetve egészen pontosan: az eredménytelenségek – vizsgálatára kijelölt bizottság szakvéleménye, az ALPAC-jelentés (Pierce et al., 1966) legtöbbjüket profilváltoztatásra kényszerítette. A korábban a georgetowni IBM-fordítókísérletet vezető magyar származású Toma Péter által alapított és az üzleti világban is sikeresnek mondható Systran rendszer az Európai Közösség érdeklődését is felkeltette, és hosszas tárgyalások után meg is vásárolták az egyre több nyelvet beszélő közösség fordítási gondjainak csökkentése céljából. A Logos fordítórendszer, melynek indulását a vietnami háború nyelvi nehézségei szolgáltatták, üzleti termékké vált, és a hetvenes évektől először a Wang, majd tőle az IBM, később pedig a Sun cég vásárolta meg, üzleti reményekkel. A Texas Egyetemen kifejlesztett angol–német fordítást végző Metal rendszer 1978-ban Európába került, a Siemenshez. A gépi fordítás az Egyesült Államokon kívül bizonyos értelemben érintetlenebb maradt az ALPAC-jelentés következményeitől. Így alakulhatott ki Kanadában az angol és francia időjárás-jelentéseket az egyik nyelvről a másikra fordító szolgáltatás, a METEO, vagy az egységes gazdaság irányába induló Európa néhány erődemonstrálási céllal indított K+F-projektje: az Eurotra és a DLT. Ez idő tájt jelentkezett az ötödik generációs számítógép gondolata is, és benne a japán álom, mely az akkor még két évtizednyi távolságban levő ezredfordulóra prognosztizálta a nyelvet intelligensen használó, beszélő és fordító számtógép megvalósítását. Mivel akkoriban

ettől még nagyon messze látszott lenni a világ, az amerikai oldalon megelégedtek az újonnan megjelenő fogalom, a *természetesnyelv-feldolgozás* (natural language processing: NLP) emlegetésével. Hazánkban egyébként az ötvenes évek végétől néhány évig szintén működött egy gépi nyelvészeti csoport, melynek kutatásait részben épp az ALPAC-jelentés közép-európai mellékhatásaként állították le (Prószéky, 2013).

A nyelvészet területén a gépi fordítás számára szóba jöhető újdonság csak a hetvenes évek végén jelentkezett, amikor Chomsky transzformációs nyelvtanának alapproblémáit egy új ügyes technikával kikerülve – bizonyos értelemben a strukturalista Harris és a generatív Chomsky közötti nyelvleírási különbségek újragondolásával – megjelent néhány új formalizmus: a GPSG, az LFG, majd a HPSG (Sells, 1985). Ezek az elképzelések azért voltak jelentősek, mert a számítógépes megvalósíthatóságot fontos szempontként maguk előtt tartva új lökést adtak a gépi nyelvészet művelőinek is. Azonban a lexikalizálódás, azaz a szótári információknak a szintaxis területén való hatékony térfoglalása meglehetősen komplex nyelvi struktúrákat és ebből következően (gép)időigényes műveletsorokat hozott. Így az ezeket működtetni szándékozó informatikai megoldások csak a fenti elméletek képességeinek demonstrálását szolgálták elsősorban, a gyakorlati életben, például a gépi fordítás területén nem játszottak meghatározó szerepet. Egy másik elméleti indíttatású gépi fordítási közelítés a modern formális logika egyik atyja, Gottlob Frege elmélete (Frege, 1923) egyfajta számítógépesítésének mondható Rosetta rendszer volt. Ez a „rule-to-rule" hipotézisen, azaz a szintaktikai és szemantikai szabályok párba állításán alapuló fordítási közelítés középpontba állításán alapult, de ennek sem lettek gyakorlati követői a gépi fordítás más művelői között. Időközben Chomsky folyamatosan megjelenő újabb generatív nyelvészeti elképzelései (Chomsky, 1981; 1993) meglehetősen átformálták a korábbi közelítést, de a generatív felfogás alapjai nem változtak, ezért a számítógépesek és különösen a gépi fordítók továbbra is jobban bíztak a hetvenes évek elején kialakult alapmodelljeikben. Ezek aktuális összefoglalását épp az a Winograd adta, aki a hetvenes évek elején bemutatott procedurális módszerével beírta magát a gépi nyelvészet történelmébe. Winograd nyelvi proceduralitásról szóló, összefoglaló, egyfajta „kvázi-formális" elméletről szóló könyve, a *Language as a Cognitive Process* 1983-ban jelent meg (Winograd, 1983). Ez idő tájt egyébként más kognitív grammatikák is megjelentek, melyek tudásalapú paradigmák formájában a gépi fordításon belül is fel-felbukkantak. Ezekben a világismeret és a nyelvi tudás keveredett, némiképp fittyet

hányva a nyelvészeti jelentéstan és a világismeret közötti falat szigorúan őrző nyelvészeti közelítéseknek. A Winograd-könyv egyik érdekessége egyébként, hogy bár összefoglalt szinte mindent, ami a számítógépes nyelvfeldolgozásban fontos lehetett a nyolcvanas évek elején, ám az a szó, hogy „morfológia", nem fordult elő benne. Itt is tetten érhető tehát, hogy a nyelv fogalma akkoriban többé-kevésbé az angol nyelvet jelentette. Ugyanebben az évben épp az említett területen történt egy fontos elméleti áttörés: a számítógépes nyelvészet morfológiai leírása egységes elméleti hátteret kapott, ugyanis megszületett egy új formalizmus, a reguláris nyelvtanok „újjászületésére" építkező kétszintes morfológia (Koskenniemi, 1983). Ettől kezdve a szabályalapú gépi fordító rendszerek legelső és legutolsó modulja, a szóalaktani elemzés és a szóalaktani generálás mostantól nem feltétlenül ad hoc karaktermanipulációkra, hanem ezekre a kétszintes rendszerekre épülhet.

A nyolcvanas évek első felében megjelentek az első személyi számítógépek, és hamarosan a számítógépes nyelvészet első piaci alkalmazásai is: a helyesírás-ellenőrző és az elválasztóprogramok (először Macintosh gépekre, majd IBM PC-re is). Nem sokkal később a gépi fordítás is megpróbált „leszállni" a személyi számítógépekre: kijött a PC Logos, majd a Siemens által megvásárolt Metal rendszer a szótárakat kiadó Langenscheidté lesz, és T1 néven – a sokak által jól ismert jellegzetes Langenscheidt-szótárak borítójához hasonló papírdobozban – a boltok kirakatába került. A Systrannak is kijött a PC-s változata, és létrejöttek az első, kimondottan a PC-s környezethez igazított képességű fordítórendszerek, mint pl. a finn Kielikone vagy az orosz ProMT. Magyarországon a nyolcvanas évek végén újra indult a számítógépes nyelvészet: megjelent az első magyar nyelvű összefoglaló az addigi eredményekről (Prószéky, 1989), majd 1991-ben létrejön az először csak nyelvhelyességi eszközöket, majd gépi fordító modulokat is létrehozó MorphoLogic cég (Mikolás, 2001).

3. A sok szöveg egyre jobbat tesz a gépi fordításnak

Miközben a PC-k hozták az első eladható gépi nyelvészeti megoldásokat, a tudomány újat lépett: beköszönt az internet és ezzel a számítógéppel távolról elérhető anyagok világa. Ráadásul egyre több anyag került ebben az időben már számítógépre, és előbb-utóbb a világhálóra is. A géppel feldolgozható szövegeknek egyfajta példatárként való használata mentén a nyelvtudománynak egy új, empirikus ága alakult ki: a korpusznyelvészet (részletesebben ld. McEnery és Hardie, 2013). Magának a

korpuszelméleti közelítésnek a gyökerei egyébként még a 19. század második felére mennek vissza, ahonnan még nagyon messze volt a számítógép. Az említett gondolatcsírák a kor egyik legnevesebb magyar nyelvészéhez, Simonyi Zsigmondhoz köthetők, akinek kis nyelvtanáról ezt olvashatjuk:

> „Simonyi új grammatikai módszert akar behozni, könyve inductive halad, azaz a példákból kiindulva tanítja a szabályt, nem pedig dogmatice. A grammaticát tehát valami olvasmány alapján akarja előadni, úgy hogy a szabályokat a tanár tanítványai közreműködésével vonhatja le ésszerű következtetések útján. Ilyenképp tehát ezen módszer véget vet a lelketlen magolásnak, és azt észfejlesztő inductióval pótolja. Eszerint a szabályok is mélyebben vésődnek be a gyermek emlékezetébe, mert amit magunk találunk, azt jobban tudjuk, mint amit más mond, vagy más tanultat velünk" (Riedl, 1882).

Erre az idézetre egyébként Sass Bálint, a Nyelvtudományi Intézet nyelvtechnológus kutatója, egykori doktoranduszom hívta fel a figyelmet, aminek lényegét mai világunkban úgy mondanánk, hogy egy új grammatikai módszer van megjelenőben, mely induktív módon halad, azaz a példákból kiindulva ismeri fel a szabályt. A grammatikát tehát az elolvasott, feldolgozott szövegek alapján építjük, úgy hogy a szabályokat a gép a példák segítségével állítja össze statisztikai következtetések útján. Ezáltal ez a módszer véget vet az előre megadott szabályok mechanikus alkalmazásának, és azt indukcióval pótolja. A szabályok így tárolódnak el a gép memóriájában, mert „amit magunk találunk, azt jobban tudjuk, mint amit más mond, vagy más tanultat velünk".

Ahol pedig megjelenik a mennyiség, ott megjelennek a valószínűségszámítási módszerek is. Így történt, hogy a kilencvenes években a statisztika „beszállt" a nyelvi modellezésbe is. A szövegek statisztikai feldolgozása ettől kezdve az IBM által kidolgozott algoritmusok alapján (Jelinek, 1997) elsősorban a beszédtechnológiából jól ismert zajoscsatorna-módszerrel történt. Ez olyan sikeresnek bizonyult, hogy rövid idő alatt kialakult a statisztikai módszerek nyelvészeti alkalmazásainak a világa. Ebben az időben jelent meg a világpiacon a belga Lernout és Hauspie, az akkoriban sikertörténetének a csúcsán járó PC-s hangkártya, a SoundBlaster két kifejlesztője. Cégük, az L&H a beszédtechnológia, sőt, a mesterséges intelligencia és a nyelvfeldolgozás rövid távú világméretű térhódítását prognosztizálta, és külső tőketámogatással elkezdték

felépíteni a terveik szerint az egész földgolyót átszövő technológiai hálózatukat, melyet SAIL-nek (= Speech, Artificial Intelligence, Language) kereszteltek el. A tervezett központok, az ún. kikötők, azaz „SAIL-portok" között még Budapest is szerepelt mint lehetséges kelet-európai központ, de az akkori magyar kormány idejében észlelte a szakmai figyelmeztetéseket, és végül nem állt be a SAIL rendszert anyagilag is támogató államok közé. A beszédfeldolgozás és a gépi fordítás L&H által ígért eredményei ugyan nagyon kecsegtetőek voltak, de az igazi és álüzletemberek hada komoly etikai, aztán jogi, majd anyagi nehézségekbe hozta az L&H vállalkozást, végül a börtönbe csukott két vállalkozó által összevásárolt nyelvtechnológiai és gépi fordító cégek hatalmas elegyét a ScanSoft, majd tőle a hazánkban a valahai Recognita karakterfelismerő cég mai tulajdonosaként ismert Nuance vásárolta meg. Érdembeli fejlesztés valójában nem sok történt az L&H környékén, de az események figyelmeztetésként hatottak sok, még éppen csak induló nyelvtechnológiai vállalkozás és az őket támogatók számára. Pozitív hozadéka volt az időszaknak, hogy a belga cég megjelenése a magyar politika legfelsőbb köreiben felhívta a figyelmet ennek az addig egyáltalán nem támogatott K+F terület létezésére. A 2000-es évek elejétől tehát hazánkban is megindultak a már központi forrásokból is támogatott nyelv- és beszédtechnológiai kutatások, és az addigra a MorphoLogic cég által kifejlesztett, angolról magyarra fordító MetaMorpho rendszer (Prószéky és Tihanyi, 2002) magyar–angol modulja már így jöhetett létre (Novák et al., 2008).

Ez volt az az időszak, amikor a világban kialakult a „human language technologies", azaz a *nyelvtechnológia* fogalma. Az IBM ezekben az években – átérezve az új kor üzenetét – komoly mesterségesintelligencia- és nyelvtechnológiai „erődemonstrációkat" tartott. Az első, a Deep Blue rendszerről szóló ugyan nem nyelvi megoldásokat, hanem a sakkozást népszerűsítette, de olyan szinten, hogy rendszerük megverte a regnáló sakkvilágbajnokot, Gari Kaszparovot (IBM, 1997). Ezzel a mesterségesintelligencia-technológiák bemutatták, hogy az alapismeretek (ez esetben a sakkfigurák lépéseinek szabályai) az eredmények szempontjából ugyan fontosak, de nem elsődlegesek, hiszen ezeket eddig is tudták a sakkprogramok, ezzel szemben rengeteg játszmát kell megfelelően elemezni és feldolgozni, mert akkor a program a sok-sok nemzetközi nagymester együttes tudásával le tud győzni gyakorlatilag akárkit, aki még ha nagyon okos is, de végül is csak egyetlen ember. A gépi fordításra alkal-

mazva ez a logika valahogy így hangzik: ha a nyelv mondatépítő szabályait ismerjük, az ugyan fontos, de ami igazán szükséges, az a rengeteg olyan minta, amit már emberek bizonyos szövegek fordításaként korábban létrehoztak. Ha a sok elérhető fordítást megtanítjuk a rendszernek, akkor a sakkprogramhoz hasonlóan fordítók ezreinek a tudását fogja tudni egyidejűleg alkalmazni (természetesen valamilyen statisztikai formában) egy adott, még le nem fordított szöveg célnyelvi megfelelőjének létrehozásához. A gépi fordításban ráadásul nem is valaki ellen kell használni ezt a tudást, mint a sakkban, hanem mindannyiunk javára. A gépi fordítás ezektől a matematikailag kifogástalan megoldásoktól tehát szárnyakra kapott, mindössze a kiinduló anyag mennyisége és minősége volt az, ami a géppel fordítandó szöveg más nyelven történő megfogalmazásának használhatóságát befolyásolta. Az új évezred első évtizedének a az IBM újabb, immár nyelvi csodarendszerként beharangozott alkalmazással állt elő, melyet a cég egyik legbefolyásosabb elnökéről, Thomas J. Watsonról neveztek el. A Watson rendszer ugyan nem a fordításban jeleskedett, hanem azt a tudást, amit az ezzel foglalkozó kutatók a rendszer számára elérhetővé tettek, viszonylag bonyolult kérdések megválaszolására használta fel. Ezt a tevékenységet természetesen fel lehet fogni úgy is, hogy a bemenő nyelvi adatot a belső keresőrendszer „nyelvére" kellett lefordítania. A rendszer demonstrációján egy népszerű kvízjáték győzteseit verte meg a televíziónézők millióinak szeme láttára (IBM, 2011).

4. A neurális hálók megjelennek a gépi fordítás területén is

Ezzel a *mesterséges intelligencia* fogalma ismét előtérbe került a nyelvfeldolgozással kapcsolatban. Nem sokkal ezután jött el az a pillanat, amikor a *mélytanulás* és a *neurális hálós módszerek* újra mesterséges intelligencia néven maguk alá gyűrték az addig kétkedő világot. Egy brnói hallgató PhD-disszertációjában kidolgozott egy olyan módszert, a *szóbeágyazást* (Mikolov, 2013), amellyel a nyelv szavait vektorokként tudta reprezentálni, méghozzá úgy, hogy a jelentésükben hasonló szavak a vektortérben közel kerültek egymáshoz, a távoliak pedig messze. Mindehhez semmilyen nyelven kívüli információt nem használt fel, mindössze a szavak különböző mondatokban talált előfordulásainak szókörnyezetét. Mivel megnyilatkozásainkban a szavak mindig mondatokban, nagyobb szövegegységekben fordulnak elő, és csak ott jelentik azt, amit, ha két szó környezete sokszor hasonló, akkor nagy eséllyel az adott szavaknak is hasonlítaniuk kell egymásra. Ez egy régóta ismert alapgondolat, hiszen tudományos megfogalmazásában ez eddig is valahogy úgy

hangzott, hogy a jel jelentése a jel használati szabálya (részletesebben ld. Wittgenstein, 1953). A jel itt a szó, és a használati szabályt a környező szavak közötti előfordulás jelenti. Mindössze az a különbség, hogy az eddigi meglehetősen absztrakt megfogalmazás helyett most Mikolov egy egzakt matematikai módszert mutatott, az ezt megvalósító programmal együtt. Ez a program a neurális hálók egyik első alkalmazása volt a nyelvtechnológiában, és alapvetően megváltoztatta a számítógépes nyelvészet világát. Az ilyen vektoros reprezentáción alapuló gépi fordító rendszerek nem a szavak betűalakját, hanem valójában ezeket a szemantikus térben megjelenő „jelentéscsomókat" fordítja, következésképp egy kicsit úgy tud viselkedni, mintha „értené" a szöveget, és nemcsak a betűit olvasná. Ez a közelítés a gépi fordítás azonnali minőségi javulását hozta. Például a gépi úton eddig nehezebben fordítható nyelvpárok minőségi ugrást mutattak, és közel kerültek azokhoz a nyelvpárokhoz, melyeket már korábban is sikeresen fordítottak a gépek. Örömünkre a magyart (és az EU más eddig nehezen kezelhető nyelveit, mint pl. a finnt vagy az észtet) tartalmazó nyelvpárok egyre használhatóbb minőségű fordításkat produkáltak. Ami viszont mind a korábbi statisztikai, mind ezeket a neurális fordítórendszereket illeti, van egy igen fontos probléma, ami a tanítóadatok mennyiségéből és minőségéből következik. Igen jelentős mennyiségű szöveg – ún. bitext, tehát forrásmondat-célmondat párokból álló kétnyelvű szövegkorpusz – szükséges a jó fordításhoz, viszont egy szűk szakterületnek még ha az összes valaha készített fordítását fel is tudnánk használni tanítóanyagnak, sokszor az is kevés a jó minőségű gépi fordításhoz. Ugyanez a probléma áll fenn azoknak a nyelvpároknak az esetében is, amelyeken az összes eddig készült fordítás együtt sem volna elég tanítóanyagnak. Gondoljuk el például a magyar–máltai gépi fordító rendszert, aminek a számára ha minden eddigi ember készített fordítást össze is szedünk, nem kapnánk megfelelő minőségű statisztikai/neurális gépi fordítást a gépi tanuláshoz a kis mennyiségűnek számító tanítóanyag miatt. Az, hogy bizonyos típusú fordítások (tehát ritka nyelvpárok, vagy gyakoribb nyelvpárok kevés fordítási mintával rendelkező szakterületei) esetében nincs megfelelő mennyiségű kiinduló anyag, a szakma „sparse data problem"-nak nevezi. Tehát mind a matematikai alapok, mind az informatikai megoldások elvileg tökéletesek, ám a nehézséget a gyakorlatban a nyelvi anyag hiánya vagy nem megfelelő minősége adja.

Ha nagyok a tanítókorpuszok, akkor viszont valószínűleg nagyon heterogének, mert mindenféle szövegtípus előfordul bennük (gondoljunk

csak az interneten fellelhető szövegek sokféleségére), így egy-egy kifejezésnek több lehetséges fordítása is előfordul bennük a különböző környezetekben. Hogy ezeket a lehetséges többértelműségeket szétválaszszuk egymástól, jó volna homogenizálni a korpuszokat, azaz szűkebb tematikus egységekre, doménekre bontani. Ezeken belül ugyanis már jóval kisebb lesz az egyes szavak többértelműsége, ám így a kiinduló korpusz mérete is kisebb lesz, ami egyfajta 22-es csapdájaként az említett „sparse data problem"-hoz vezethet. Előáll tehát a statisztikai/neurális rendszereknek egy nehezen feloldható kettőssége: ha kicsi a szövegkorpusz, bár a tanítóminta ilyenkor nagyrészt egyértelmű szavakat tartalmaz, sokszor nem lesz jó az erre épülő fordítás az egyes kifejezések relatíve kis előfordulási száma miatt. Ha növeljük a korpusz méretét, óhatatlanul megjelenik a többértelműség okozta „fordítási zaj", bár a korpusz mérete már más szempontból megfelelőnek tűnhet.

Egy másik nagy probléma napjaink neurális gépi fordításában, hogy az informatikai kutatóközpontokban ugyan készülnek nyelv(pár)független modellek, ám ezek minősége meg sem közelíti a nyelv(pár)specifikus modellekét. Nyilván nem minden kutatóhely rendelkezik minden nyelvre megfelelő mennyiségű olyan tanítóanyaggal, amiből jó minőségű fordítás volna várható. Ráadásul a neurális rendszerek nyelvmérnökei elsősorban nem a fordításban jók, de még csak nem is abban, hogy előkészítik a nyelvi anyagokat a programrendszerek számára, hanem abban, hogy a neurális megoldáshoz szükséges felfoghatatlan mennyiségű paramétert úgy állítják be, hogy a fordítóprogram jó minőségű eredményt adjon. A paraméterbeállítások mikéntje viszont jelenlegi tudásunk szerint nehezen hozható közvetlen logikai kapcsolatba az eredménnyel, tehát a gépi nyelvészet világában mindig is jelen lévő intuíciónak még jobban felértékelődik a szerepe a mai gépi fordító rendszerek létrehozásánál. Ha egy nagyobb cégnél sok intuitív ember jön össze, és ezeken a helyeken a géppark lehetőségei is komoly sebességelőnyt mutatnak egy kisvállalkozás gépeivel szemben, hamar megérthetjük, hogy ugyanolyan intuitív emberek kisebb kapacitású gépekkel nagyságrendekkel kevesebb kísérletet tudnak végezni ugyanannyi idő alatt a paraméterbeállítások világában, mint nagy céges társaik. Tehát a gépi fordítás területén a verseny ma elsősorban nem a nyelvi vagy programozási tudáson múlik, hanem a kísérletezésen, amelyben a gyorsabb környezet előbb jelzi vissza egy-egy kísérlet eredménytelenségét, mint a lassabbé. És ha mindezt több tízszer vagy százszor annyi kísérletező ember végzi, hamar belátható, hogy né-

hány világcég jelentős előnyt tud szerezni a mai gépi fordítási versenyben, mint bármikor korábban. Más szavakkal: egyre jobban nyílik az olló a kis és a nagy gépi fordító intézmények között. Egy dolog ugyanakkor egyre jobban látszik: az általános modellek általában nem elegendőek egy adott nyelvi közösség számára, hiszen az ilyen modellek azért készülnek, hogy relatíve kevés munkával lehessen összehasonlítható eredményeket felmutatni a bármely nyelvről bármely nyelvre való fordítás világában. Akiknek viszont az általános nyelvtechnológiai eszközök világában egy konkrét nyelvre, vagy a gépi fordítás esetében egy-egy konkrét nyelvpárra kell egyre jobb eredmény, azoknak a saját tanítóadataik egyre jobb minőségén és egyre nagyobb mennyiségén kell dolgozniuk, még ha az ezeket feldolgozó szoftverek mindössze néhány világcég műhelyéből jönnek is elő. És ebből következően talán mindenki számára érthető, hogy a neurális megoldások világában is van értelme támogatni a magyar nyelvtechnológiai fejlesztéseket, és ezáltal a magyarról és a magyarra történő gépi fordítást is, mert helyettünk ezt mások nem fogják jó minőségben megcsinálni.

5. A gépi fordítás és Váradi Tamás találkozásai

Ami Váradi Tamásnak a gépi fordítás világához való szakmai hozzájárulását illeti, személyes tapasztalataim is vannak, mert korábban több alkalommal is összetalálkoztunk a fenti kutatások egyikét-másikát megvalósító projektekben mint partnerek. Akkor még javában külön intézményekben dolgoztunk: Tamás az MTA Nyelvtudományi Intézet, én pedig a MorphoLogic képviseletében vettem részt gépi fordítás témájú munkálatokban. Az utóbbi négy évben azonban már intézményesen is egy hajóban ülünk, és – amint ezt mindjárt prognosztizálom is – könnyen elképzelhető egy közelgő, immár belső, újabb együttműködés is. Az első magyarországi gépi fordítási projekt egyébként a 2000 és 2002 között futó MATCHPAD (Machine Translation Systems for the Use of Hungarian and Polish Administrations) volt, amikor a korábbiakban már említett Systran rendszer magyarra és lengyelre való alkalmazhatóságának bizonyítása volt a cél. Ennek a kísérleti kutatásnak az idején már bontogatta szárnyait Tihanyi László szakmai vezetésével MorphoLogic cég MetaMorpho nevű gépi fordító rendszere (Novák et al., 2008), amelynek első publikus bemutatója 2001-ben, a MorphoLogic tizedik születésnapján volt. Néhány év múlva, miután a MetaMorpho angol–magyar modulja teljesen elkészült, a MorphoLogic megcélozta a magyar–angol változatot is, amihez immár külső partnerek is csatlakoztak: az MTA NYTI

és a SZTE (Tihanyi, 2007). Így újra együttműködhettünk egy Tamás vezette csapattal, amihez később még hozzákapcsolódott a MorphoLogic webforditas.hu portáljának alapgondolatából kinövő, több európai gépi-fordítás-szolgáltatót egy nagy nemzetközi fordítórendszerré összekapcsoló iTranslate4.eu projekt Tamás általi menedzselése is. Legutóbbi összetalálkozásunk területe, amit csak futtában említettem néhány sorral feljebb, a mai nyelvtechnológia javarészt mélytanulásos technológiákon alapuló megközelítése, vagyis amit ma a sajtó – némiképp összemosva a részleteket – mesterségesintelligencia-alapúnak nevez. Ebben, különösen a legújabb transzformer rendszerek létrehozásában Tamás már rövid idő alatt is sok eredményt ért el, így nem lehetetlen, hogy előbb-utóbb a vezetésével elkészült neurális modellek a gépi fordítás területén is bizonyíthatnak. Mindehhez jó tudni, hogy a saját intézeti eredményeinken túl azokból a szakmai közösségekből, melyek együttműködéséről az előbb szóltam, a MorphoLogic MetaMorpho projektjét vezető Tihanyi László és Tamás kutatócsoportjának egyik korábbi oszlopa, Oravecz Csaba ma az Európai Bizottság Fordítási Főigazgatóságának elismert kutatói, akik ezen a területen elért eredményeikkel, azaz neurális gépi fordítási megoldásaikkal bevezették a magyar nyelvet a világ fordítóprogramjainak használható forrás- és célnyelvei közé.

Egy kutatónak, aki régóta vezethet másokat, kétféle nagyszerű szakmai eredmény létezik: ha maga ér el új eredményeket, ahogy Tamás a legújabb típusú nyelvmodellek, a magyar BERT-Large létrehozásában (Feldmann et al., 2021), valamint ha az általa hosszú ideig menedzselt kutatócsapat valamely tagja maga is elér ilyeneket (Tihanyi és Oravecz, 2017). Kedves Tamás, kívánom, hogy mindkettőből a továbbiakban is sok jusson Neked!

Bibliográfia

Bar-Hillel, Y.: The Present State of Research on Mechanical Translation. American Documentation 2/4, 229–237 (1951)

Chomsky, N.: Syntactic Structures. Mouton (1957)

Chomsky, N.: Aspects of the Theory of Syntax. MIT Press (1965)

Chomsky, N.: Lectures on Government and Binding. Foris (1981)

Chomsky, N.: A Minimalist Program for Linguistic Theory. MIT Occasional Papers in Linguistics No. 1. Cambridge: MIT Press (1993)

Feldmann Á., Hajdu R., Indig B., Sass B., Makrai M., Mittelholcz I., Halász D., Yang Zijian Gy., Váradi T.: HILBERT, magyar nyelvű BERT-large modell tanítása felhő környezetben. In: Berend G., Gosztolya G., Vincze V.(szerk.) XVII. Magyar Számítógépes Nyelvészeti Konferencia. 29–36. Szegedi Tudományegyetem TTIK, Szeged (2021)
Frege, G.: Logische Untersuchungen. Dritter Teil: Gedankenfuge. In: Beiträge zur Philosophie des Deutschen Idealismus, Vol. III. pp. 36–51 (1923)
Harris, Z. S.: Methods in Structural Linguistics. University of Chicago Press, Chicago (1951)
Hutchins, J.: From First Conception to First Demonstration: the Nascent Years of Machine Translation, 1947–1954. In: A Chronology. Machine Translation 12/3, 195–252 (1997)
701 Translator, IBM Press Release, January 8, 1954. (1954)
Deep Blue Accepts Challenge to Compete in Ultimate Chess Match with Human Champ Kasparov. IBM Press Release, May 30, 1995. (1995)
Jeopardy! And IBM Announce Charities to Benefit from Watson Competition. IBM Press Release, Jan 13, 2011. (2011)
Jelinek, F.: Statistical Methods for Speech Recognition. MIT Press, Cambridge (1997)
Koskenniemi, K.: Two-level Morphology: A General Computational Model for Word-Form Recognition and Production. Publications, No. 11, University of Helsinki, Department of General Linguistics, Helsinki (1983)
McEnery, T.; Hardie, A.: Corpus Linguistics: Method, Theory and Practice. Cambridge University Press, Cambridge (2012)
Mikolás Z. (szerk.) MetaMorpho: A MorphoLogic tíz éve. Budapest: MorphoLogic (2001)
Mikolov, T.: Statistical Language Models Based On Neural Networks. Ph.D. Thesis: Masaryk University, Brno (2013)
Novák A., Tihanyi L., Prószéky G.: The MetaMorpho Translation System. In: Callison-Burch, C., Koehn, P., Monz, C., Schroeder, J. Shaw Fordyce, C. (eds.) Proceedings of the Third Workshop on Statistical Machine Translation at ACL. pp. 111–114. Association for Computational Linguistics, Stroudsburg, PA (2008)
Papp F.: Matematikai nyelvészet és gépi fordítás a Szovjetunióban. OMKDK, Budapest (1964)
Pierce, J. R., Carroll, J. B. et al.: Language and Machines – Computers in Translation and Linguistics. ALPAC Report, National Academy of Sciences, National Research Council, Washington, DC (1966)
Prószéky G.: Számítógépes nyelvészet: Természetes nyelvek használata számítógépes rendszerekben. Számalk, Budapest (1989)
Prószéky G., Tihanyi L.: MetaMorpho: A Pattern-Based Machine Translation System. In: Proceedings of the 24th 'Translating and the Computer' Conference. pp. 19–24. ASLIB, London, United Kingdom (2002)
Prószéky G.: A magyar számítógépes nyelvészet történeti áttekintése. In: Prószéky G., Váradi T. (szerk.) Általános Nyelvészeti Tanulmányok XXIV: Nyelvtechnológiai kutatások. pp. 17–45. Akadémiai Kiadó, Budapest (2012)
Riedl F.: Simonyi kis nyelvtana. Egyetemes Philologiai Közlöny 6/6, 573–590 (1882)
Simonyi Zs.: Kis magyar nyelvtan mondattani alapon. Negyedik átdolgozott s gyakorlatokkal bővített kiadás egy kötetben (1882)

Sells, Peter. Lectures on Contemporary Syntax Theories: An Introduction to Government-Binding Theory, Generalized Phrase Structure Grammar, and Lexical-Functional Grammar. CSLI, Stanford (1985)
Senellart, J., Dienes, P., Váradi, T.: New Generation Systran Translation System. In: Senellart, J., Yang, J., Rebollo, A. (eds.) Proceedings of MT Summit VIII. Santiago de Compostela, Spain (2001)
Tesnière, Lucien: Éléments de syntaxe structurale. Libraire C. Klincksieck, Paris (1959)
Tihanyi L.: A MetaMorpho projekt 2007-ben – a sorozat vége. In: Tanács A., Csendes D.(szerk.) V. Magyar Számítógépes Nyelvészeti Konferencia. pp. 179–186. SZTE, Szeged (2007)
Tihanyi L., Oravecz Cs.: First Experiments And Results in English-Hungarian Neural Machine Translation. In: Vincze V. (szerk.) XIII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 275–286. SZTE, Szeged (2017)
Winograd, T.: Understanding Natural Language. Cognitive Psychology 3/1, 191 (1972)
Winograd, T.: Language as a Cognitive Process. Vol. 1. Syntax. Reading: Addison-Wesley (1983)
Wittgenstein, L.: Philosophical Investigations. Blackwell, Oxford (1953). [Magyar fordítás: Filozófiai vizsgálódások. Atlantisz, Budapest (1992), ford.: Neumer Katalin]
Woods, W. A.: Transition Network Grammars for Natural Language Analysis. Communications of the ACM 13/10, 591–606 (1970),
Woods, W. A.: Progress in Natural Language Understanding: An Application to LUNAR Geology. Proceedings of the National Computer Conference AFIPS pp. 441–450 (1973)