

## Intelligens fordítómemória fejlesztése

Magyar nyelvű szakmai összefoglaló

Az internet, illetve az információ – a szöveg – továbbításának globalizációja felerősítette a fordítás iránti igényt. A XXI. század elején egységnyi idő alatt sokkal több szöveg keletkezik, mint korábban, ráadásul ezek a szövegek pillanatok alatt megjelennek idegen nyelvű környezetben is. Magyarország esetében a fordítási feladatot növeli a küszöbön álló EU-csatlakozás is. Az utóbbi évtizedben már magától értetődőnek számít, hogy a szakmai – nem szépirodalmi – fordításokhoz segítségül kell hívni a számítógépet: a feladat különben nem végezhető el a rendelkezésre álló idő alatt.

Az automatikus gépi fordítás kifejlesztésének nehézségei azonban óvatossá tették a fejlesztőket. Ez azt jelenti, hogy a teljesen automatikus gépi fordítás helyett a fordítási részfeladatokat próbálták számítógépesíteni. Rövid idő alatt így számos olyan eszköz jelent meg a piacon, amelyek révén a számítógép támogatja a humán fordító munkáját.

Pályázatunkban a géppel támogatott emberi fordítás (GTEF) legfontosabb eszközének, a *fordítómemóriának* a továbbfejlesztésével foglalkozunk. A fordítómemória alapfeladata, hogy a *memóriájában* – adatbázisában – tárolt régi forrásszövegekben megtalálja azokat a mondatokat, esetleg mondatnál kisebb kifejezéseket, amelyek leginkább hasonlítanak az éppen fordítandó mondathoz vagy kifejezéshez. Ha van az adatbázisban olyan mondat, amely eléggé hasonlít az aktuális fordítanivalóhoz, a fordítómemória felajánlhatja az adatbázisbeli mondat adatbázisbeli fordítását mint fordítási javaslatot.

A kereskedelemben számos fordítómemória-program kapható (Trados TWB, ATRIL Déjà Vu, hogy csak a legsikeresebbeket említsük) – magyar termék nincs közöttük. Mivel ezek természetüknél fogva nyelvfüggetlenek, értelmetlen a magyar nyelvre való felkészítésükről beszélni. Azonban az ilyen programok legfőbb hibája épp nyelvfüggetlenségükben rejlik: a mondatok hasonlóságát a karakterkódok matematikai vizsgálatára építik. Mivel a karakterkódok önmagukban nem hordoznak nyelvi információt, a jelenlegi rendszerekben alkalmazott eljárás minden szempontból inadekvát, alkalmazását egyedül a fejlesztő cégek gazdasági szempontjai indokolják.

A nyelvi intelligencia alkalmazása – amikor nemcsak a mondatokat alkotó karaktorsorozatokat, hanem például a nyelvtani elemzéssel kapott, bizonyos fokig absztrakt struktúrákat hasonlítjuk össze – pontosabb hasonlósági keresést tesz lehetővé, s a matematikai eljárásokhoz képest távolabbi, ám nyelvi szempontból adekvát hasonlóságokat is fel tud ismerni. Az ilyen megoldás azonban óhatatlanul nyelvfüggő lesz, mert képesnek kell lennie a forrásnyelv mondatainak nyelvi elemzésére. Mivel kezdetben az ilyen termékekben rendelkezésre álló nyelvek készlete meglehetősen szűk (fejlesztésünk, bár az alapmechanizmusokban nyelvfüggetlenségre törekszik, kezdetben az angol és a magyar nyelv feldolgozására irányul), lényegesen kisebb az elérhető piac is, s a fejlesztésbe fektetett összeg is lassabban térül meg.

A pályázat a fenti fordítómemória-technológia, illetve a rá épülő termék prototípusának kidolgozására irányul. A rendelkezésre álló két évben mind a nyelvi/nyelvtani absztrakció, mind pedig az eddigieknél hatékonyabb matematikai módszerek lehetőségeit megvizsgáljuk, s a két eljárástípust együtt implementáljuk. Az eredményül kapott technológia így rendelkezik majd a kereskedelmi forgalomban kapható fordítómemóriák előnyeivel is, és megfelelő nyelvi adatbázis híján nyelvfüggetlen fordítómemóriaként is használható.

A kereskedelmi fordítómemóriák nagy hátránya az is, hogy üresen kerülnek a felhasználóhoz. Projektünkben kifejlesztünk egy kétnyelvű, informatikai szövegekből álló, nyelvenként körülbelül 1 millió szavas, géppel annotált referenciakorpuszt is, amelyet egyfelől a nyelvi absztrakció vizsgálatára használunk, másfelől kialakítunk belőle egy olyan kezdeti adatbázist, amelynek segítségével az „alapszintű” fordítómemória azonnal képes számos nyelvi alakulat felismerésére. A referenciakorpuszt kutatási célokra hozzáférhetővé tesszük.

Bár a pályázat nyomán kifejlesztendő rendszer kezdetben idegen-magyar, illetve magyar-idegen nyelvpárokat alkalmaz, nem szabad elfelejteni, hogy fejlesztésünk – jelenlegi tudásunk szerint – világviszonylatban új technológia létrehozását célozza. A rendszer további nyelvpárokra való alkalmazása és külföldi megismertetése növelné a magyarországi kutatás-fejlesztés presztízsét.

A projekttel kapcsolatos támogatási igényt az indokolja, hogy a nyelvfüggő nyelvi intelligenciát tartalmazó programok fejlesztése lényegesen nagyobb tőkét igényel, és a programok is sokkal lassabban terjeszthetők el, mint nyelvfüggetlen társaik. A várható minőségjavulás és a fordítások csökkenő munkaigénye azonban mindenképpen szükségessé teszi a fejlesztést. A fejlesztéshez szükséges tőke nem áll teljes egészében az iparág rendelkezésére, amit az is mutat, hogy a fordítómemóriákat fejlesztő és forgalmazó nemzetközi cégek sem költenek rá.

A későbbiekből látható, hogy a fejlesztésre egy, az IKTA-projektek történetében szokatlan konzorcium vállalkozik. A MorphoLogic Kft. régóta fejleszt fordítástámogató eszközöket, számára ez a fejlesztés a továbblépés szerves része. A partner pedig – a SZAK Kiadó – olyan könyvkiadó vállalat, amely kezdettől fogva igyekszik szisztematikusan kezelni a fordítási projekteket, ugyanis a legtöbb esetben rendkívül rövid idő alatt kell megjelentetnie nagy terjedelmű könyvek fordításait. A kiadónál jelenleg előkészítés alatt áll az eddig fordított szövegekből létrehozandó párhuzamos szaknyelvi korpusz. A kiadó fordítási tapasztalatai megfelelő tudásközponttá teszik a projektben, a fordítási projektek pedig ideális tesztkörnyezetet kínálnak a készülő eszköz számára.

Bár a pályázatban vázolt kutatási feladat számos alapkutatói lépést is igényel, jelen tudásunk szerint Magyarországon nincs olyan közfinanszírozású kutatóhely, amelyben ez a fajta kutatás hasonló fokon elő lenne készítve. Ezért döntöttünk úgy, hogy a pályázatot – az alapkutatáshoz járó támogatás elvesztése árán is – ebben a konzorciumban nyújtjuk be.

A projekt eredményeként létrejövő fordítómemória-technológia azonnal alkalmazható angol-magyar és magyar-angol fordítási feladatok elvégzésében. A kétéves fejlesztési folyamat lezárulása után mód nyílik további nyelvek gyors hozzáadására is, mivel a nyelvi alatechnológia kialakításakor nyelvfüggetlenségre törekszünk. Várakozásunk szerint a technológia lényegesen felgyorsítja a humán fordítást, és javítja annak minőségét is; véleményünk – s jelen tudásunk – szerint ez a projekt nemcsak Magyarországon hiánypótló, hanem pótolja a teljes iparág egy régi adósságát is.