Validacija označevanja diskurznih označevalcev v korpusih Turdis-2 in BNSIint

Darinka Verdonik[1], Andrej Žgank[1], Agnes Pisanski Peterlin[2]

[1]Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru
Smetanova 17, SI-2000 Maribor
{darinka.verdonik, andrej.zgank}@uni-mb.si
[2]Filozofska fakulteta Univerze v Ljubljani
Aškerčeva 2, SI-1000 Ljubljana
agnes.pisanski@guest.arnes.si

Povzetek

Označevanje diskurznih označevalcev v korpusnem gradivu je lahko včasih odvisno od interpretacije označevalca. Da bi ocenili, koliko so rezultati korpusne analize diskurznih označevalcev odvisni od interpretacije označevalca korpusnega gradiva in natančnost uporabljene sheme za označevanje diskurznih označevalcev v slovenščini, smo izvedli validacijo označenosti reprezentativnega vzorca uporabljenega korpusnega gradiva. Rezultati so pokazali, pri katerih diskurznih označevalcih se pojavlja večja variabilnost označevanja in s katerimi diskurznimi označevalci bi bilo mogoče shemo nadgraditi.

Validating the annotation of discourse markers in Turdis-2 and BNSIint corpora

The annotation of discourse markers in a corpus may sometimes depend on annotator interpretation. To assess to what extent the results of a corpus analysis of discourse marker use depends on annotator interpretation and to evaluate the precision of the annotation scheme used in the annotation of discourse markers in Slovene, a validation of the annotation of a representative sample of the corpus material used was carried out. The results showed which discourse markers show greater variability and which discourse markers could be used to upgrade the annotation scheme.

1. Uvod

Diskurzni označevalci so zadnji dve desetletji v pragmatičnem jezikoslovju zelo aktualna tema (Schiffrin, 1987; Redeker, 1990; Fraser, 1999; Schourup, 1999; Blakemore, 2002; Fox Tree, 2006; idr.; v slovenističnem jezikoslovju pa npr. Gorjanc, 1998; Smolej, 2004; Verdonik, 2006; Pisanski Peterlin, 2005; Schlamberger Brezar, 2007; Verdonik et al., 2007a) in v splošnem pomenijo številne predvsem pragmatične izraze, ki v diskurzu ne prispevajo (pomembno) k vsebini, kot npr. v naslednjem segmentu odgovora informatorke v turistični agenciji stranki (diskurzni označevalci označeni s krepkim tiskom):

zdaj pa *zlo pomembno kar je | tudi pri* ***eee*** *pri* ***eee*** *tejle kar se vstopnici tiče* ***ne?***

Na področju jezikovnih tehnologij najdemo vedno več poskusov označevanja diskurznih označevalcev v jezikovnih virih (Carlson et al., 2003; Mitkov et al., 2000; Muller et al., 2002; Byron et al., 1997; Heeman, Allen, 1999; Miltsakaki et al., 2002). Tovrstne poskuse spodbuja predvsem potreba po dodajanju vse več metajezikovnih podatkov v jezikovne vire, ki izhaja iz razvoja zahtevnejših jezikovnotehnoloških aplikacij, kot so razpoznavanje spontanega govora, strojno prevajanje, prevajanje govora, zahtevnejši sistemi dialoga ipd. Na primeru slovenskega jezika je bila shema za označevanje diskurznih označevalcev predstavljena v Verdonik et al. (2007a), na njeni podlagi pa so bili diskurzni označevalci označeni v dveh govornih korpusih omejenega obsega (vsak okoli 30.000 pojavnic), Turdis-2 in BNSIint (Verdonik et al., 2007b; Verdonik et al., v tisku). Toda označevanje diskurznih označevalcev je v veliki meri odvisno od interpretacije označevalca: isti izrazi lahko namreč opravljajo ali funkcijo diskurznega označevalca ali propozicijske vsebine, vloge pa niso vedno jasno razmejene; poleg tega nabor izrazov v vlogi diskurznih označevalcev v Verdonik et al. (2007a) nikakor ni končen. Zato lahko domnevamo, da bi različne osebe lahko bolj ali manj različno interpretirale in označevale diskurzne označevalce v korpusnem gradivu. Temu se seveda želimo kolikor mogoče izogniti oziroma moramo to upoštevati pri rezultatih korpusne analize.

Ker smo korpusa Turdis-2 in BNSIint uporabljali za jezikoslovne raziskave diskurznih označevalcev in ker želimo pred označevanjem obsežnejšega gradiva zagotoviti kar se da homogeno označevanje, smo validirali označenost navedenih korpusov[1]. Namen validacije je bil dvojen: (1) oceniti, do kolikšne mere so rezultati korpusne analize diskurznih označevalcev, ki smo jih uporabljali v jezikoslovnih raziskavah, odvisni od interpretacije označevalca korpusa; ker smo v raziskavah uporabljali samo skupne kvantitativne podatke (Verdonik et al., 2007b; Verdonik et al., v tisku), nas je tudi pri validaciji zanimala predvsem skupna kvantitativna razlika v številu označenih diskurznih označevalcev; (2) preliminarno oceniti, ali je shema za označevanje diskurznih označevalcev, predstavljena v Verdonik et al. (2007a), dovolj natančna ali pa jo je treba dopolniti in v katerih segmentih.

V nadaljevanju najprej predstavimo oba validirana korpusa, Turdis-2 in BNSIint, nato opišemo validacijski postopek in predstavimo rezultate validacije.

2. Gradivo

Korpusa Turdis-2 in BNSIint zajemata vsak približno 30.000 pojavnic.

Turdis-2 vključuje telefonske pogovore med stranko in informatorjem v turistični agenciji, turistični pisarni

[1] Delo enega izmed soavtorjev je bilo delno sofinancirano s strani ARRS po pogodbi št. J2-9742-0796-06.

oziroma hotelski recepciji. Gradivo je izbrano iz korpusa *Turdis* (Verdonik, Rojc, 2006) tako, da obsega okoli 30.000 pojavnic. Na tak obseg smo se omejili zaradi vzporednih jezikoslovnih raziskav (Verdonik et al., 2007b; Verdonik et al., v tisku), v katere smo zajeli gradivo dveh različnih pogovornih žanrov v primerljivem obsegu. 30.000 pojavnic je dvakrat več gradiva kot v naših predhodnih raziskavah (Verdonik, 2006; Verdonik et al., 2007a). Gradivo je omejeno zaradi časovne zahtevnosti ročnega označevanja korpusov in razpoložljivosti primernih govornih virov.

Gradivo za korpus *Turdis* je bilo posneto na Fakulteti za elektrotehniko, računalništvo in informatiko v Mariboru spomladi 2004 v sodelovanju z lokalnimi turističnimi organizacijami in njihovimi zaposlenimi. Korpus je bil ročno ortografsko transkribiran. Natančnejši podatki o izboru *Turdis-2* so predstavljeni v tabeli 1.

	Št. pog.	**Povprečna dolžina**		**Skupna dolžina**	
		Minute	Pojavnice	Minute	Pojavnice
Turistična agencija	38	3,40	525	129,23	19936
Turistična pisarna	12	3,63	529	43,58	6350
Hotelska recepcija	15	2,78	417	41,68	6261
Skupaj	**65**	**3,30**	**501**	**214,49**	**32547**

Tabela 1: Število in dolžina pogovorov v Turdis-2.

Korpus BNSIint vključuje televizijske intervjuje o aktualnih dogodkih v večerni dnevnoinformativni oddaji nacionalne televizije iz obdobja 1999-2005. V intervjujih sodelujejo novinar ter en ali dva intervjuvanca. Gradivo je izbrano iz baze *BNSI Broadcast News* (Žgank et al., 2004), ki je nastajala v sodelovanju Fakultete za elektrotehniko, računalništvo in informatiko v Mariboru ter RTV Slovenija in vključuje dnevnoinformativne oddaje in informativne pogovorne oddaje, zajete iz arhiva RTV Slovenija. Oddaje so bile ročno ortografsko transkribirane in segmentirane. Natančnejši podatki o korpusu BNSIint so v tabeli 2.

Št. pog.	**Povprečna dolžina**		**Skupna dolžina**	
	Minute	Pojavnice	Minute	Pojavnice
30	**6,61**	**1041**	**198,35**	**31236**

Tabela 2: Število in dolžina intervjujev v BNSIint.

3. Validacija

Preverjanje kakovosti je bistven element gradnje jezikovnih virov. Naš namen je bil validirati kakovost označevanja diskurznih označevalcev v Turdis-2 in BNSIint. V ta namen smo pripravili validacijski korpus; ta je obsegal približno 10 % gradiva iz korpusov Turdis-2 in BNSIint. Podrobnejši podatki o obsegu validacijskega korpusa (število pogovorov, dolžina v minutah, število pojavnic) so predstavljeni v tabelah 3 (za Turdis-2) in 4 (za BNSIint).

	Št. pogovorov		**Dolžina v min.**		**Št. pojavnic**	
	Skupaj	%	Skupaj	%	Skupaj	%
Turistična agencija	4	10,5	13,78	10,7	1965	9,6
Turistična pisarna	2	17,6	4,97	11,4	638	10,1
Hotelska recepcija	2	13,3	3,88	9,3	578	9,2
Skupaj	**8**	**12,3**	**22,63**	**10,6**	**3181**	**9,8**

Tabela 3: Število pogovorov ter dolžina v minutah in številu pojavnic validacijskega korpusa skupno ter v odstotkih od celotnega gradiva Turdis-2.

Št. pogovorov		**Dolžina v min.**		**Št. pojavnic**	
Skupaj	%	Skupaj	%	Skupaj	%
3	**10,0**	**18,62**	**9,39**	**3157**	**10,11**

Tabela 4: Število pogovorov ter dolžina v minutah in številu pojavnic validacijskega korpusa skupno ter v odstotkih od celotnega gradiva BNSIint.

V korpusih Turdis-2 in BNSIint so bili diskurzni označevalci ročno označeni skladno s shemo, predstavljeno v Verdonik et al. (2007a). Validacija je potekala tako, da sta diskurzne označevalce v validacijskem gradivu na novo označila dva zunanja strokovnjaka, ki nista bila povezana s snovanjem sheme za označevanje (Verdonik et al., 2007a) in označevanjem obeh korpusov, ju pa označevanje diskurznih označevalcev v korpusih zanima z različnih uporabnostnih vidikov: prvi validator je bil namreč strokovnjak s področja uporabnega jezikoslovja, drugi s področja govornih tehnologij.

Validacijsko označevanje diskurznih označevalcev je bilo opravljeno skladno s shemo za označevanje, ki je predstavljena v Verdonik et al. (2007a). V tej shemi so diskurzni označevalci definirani kot izrazi, ki k vsebini diskurza ne prispevajo nič ali skoraj nič, pojavljajo pa se v naslednjih pragmatičnih funkcijah:

- vzpostavljanje povezave z vsebino prejšnjega oziroma sledečega diskurza,
- vzpostavljanje in razvijanje odnosa med sogovorniki,
- izražanje odnosa govorca do prejšnje oziroma sledeče vsebine diskurza,
- organiziranje poteka diskurza na ravni prehodov med temami pogovora, menjavanja vlog in strukture izjave.

V Verdonik et al. (2007a) so nadalje najpogostejši tovrstni izrazi tudi našteti in obravnavani, in sicer so to: *ja, mhm, aha, aja, no, eee* v različnih izgovornih variantah (*eeem, eeen, nnn, mmm ...*), *ne?/a ne?/ali ne?/jel?, dobro/v redu/okej/prav, glejte/poglejte, veste/a veste/veste* + vprašalni zaimek (npr. *veste kaj*), *mislim, zdaj* in oporni signali. Slednji zaradi načina transkripcije gradiva niso mogli biti validirani.

	Turdis-2						BNSIint					
	K	V1		V2		E_t%	K	V1		V2		E_b%
DO	***F***	***F1***	***E1 %***	***F2***	***E2 %***		***F***	***F 1***	***E1%***	***F 2***	***E2 %***	
glejte	5	6	-	5	-	**-**	15	14	+6,7	15	0,0	**±3,4**
ja	66	64	+3,0	70	-6,1	**±4,6**	13	13	0,0	13	0,0	**±0,0**
ne?	65	62	+4,6	65	0,0	**±2,3**	16	11	+31,3	18	-12,5	**±21,9**
dobro idr.	36	36	0,0	36	0,0	**±0,0**	5	5	-	5	-	**-**
mislim	4	2	-	4	-	**-**	1	1	-	3	-	**-**
zdaj	16	15	+6,3	20	-25,0	**±15,7**	1	1	-	1	-	**-**
SKUPAJ	**192**	**185**	**+3,6**	**200**	**-4,2**	**±3,9**	**51**	**45**	**+11,8**	**55**	**-7,8**	**±9,8**

Tabela 5: Rezultati validacijskega označevanja in vrednotenja označenosti gradiva.

Validatorja se o odprtih vprašanjih nista smela posvetovati ne med seboj ne z avtorji sheme za označevanje.

Po končanem validacijskem označevanju smo primerjali označenost validacijskega gradiva v korpusih Turdis-2 in BNSIint z označenostjo validacijskega gradiva pri prvem in drugem validatorju ter na podlagi tega ocenili, pri katerih izrazih je označevanje diskurznih označevalcev najbolj variiralo. Ker smo v raziskavah uporabljali samo skupne kvantitativne podatke (Verdonik et al., 2007b; Verdonik et al., v tisku), nas je tudi pri validaciji zanimala samo skupna kvantitativna razlika v številu označenih diskurznih označevalcev in smo opazovali le skupno razliko v številu označenih diskurznih označevalcev, ne pa tudi razlik pri označevanju posameznih pojavnic. Rezultati so predstavljeni v nadaljevanju.

4. Rezultati

Kot ugotavljajo Verdonik et al. (2007a), so nekateri izrazi vedno v vlogi diskurznega označevalca. Validatorja sta se strinjala, da so takšni izrazi *aha, aja, mhm, no* in *eee/eeem/eeen/nnn/mmm* ipd. Te lahko zato avtomatsko označimo in nadaljnja validacija označevanja ni smiselna.

Rezultati validacije za ostale diskurzne označevalce, obravnavane v Verdonik et al. (2007a), so prikazani v tabeli 5. Na levi strani tabele so podatki za korpus Turdis-2, na desni za korpus BNSIint. V stolpcih K je število diskurznih označevalcev v validiranih korpusih, v stolpcih V1 so podatki za gradivo validatorja 1 ter v stolpcu V2 za gradivo validatorja 2. Za vsako validacijsko gradivo je v stolpcih F1 in F2 navedena pogostost rabe. V stolpcih E1 % in E2 % je ovrednoteno odstopanje po enačbi E1 = (F – F1) / F * 100 oz. E2 = (F – F2) / F * 100. Odstopanje je izračunano tudi skupno za vsak korpus po enačbi E_t = (| E1| + |E2|) / 2 oz. E_b = (|E1| + |E2|) / 2.

Če je bilo v validacijskem gradivu manj kot 10 primerov rabe posameznega diskurznega označevalca, odstopanj v odstotkih nismo računali. To se je dogajalo predvsem pri tistih diskurznih označevalcih, ki tudi v celotnih korpusih Turdis-2 in BNSIint niso bili rabljeni pogosto (manj kot 20-krat sta npr. rabljena *mislim* v obeh korpusih in *zdaj* v BNSIint).

Kot lahko sklepamo na podlagi podatkov v tabeli 5, je najbolj nedvoumno označevanje diskurznih označevalcev *dobro/v redu/okej/prav*. Pri teh v rezultatih ni bilo odstopanj. Odstopanja do 5 %, kar je običajna tolerančna meja v validaciji, zasledimo pri diskurznih označevalcih *ja* in *glejte*. Pomembna odstopanja pa se pojavijo pri označevanju *zdaj* v Turdis-2 (v BNSIint je *zdaj* rabljen le trikrat v celotnem korpusu, zato se tam ne pokažejo pomembnejše razlike) ter pri označevanju *ne?* v BNSIint. Deloma lahko visok odstotek odstopanja pri slednjih pripišemo nizki pogostosti, vseeno pa to kaže, da je označevanju teh dveh diskurznih označevalcev treba posvetiti več pozornosti.

Medtem ko lahko za nabor izrazov, ki so že v Verdonik et al. (2007a) definirani kot diskurzni označevalci z veliko pogostostjo, ugotavljamo v splošnem dokaj homogeno označevanje tako med validatorjema kot v izvornih korpusih, pa so velike razlike pri označevanju ostalih, "novih" izrazov, ki po presoji validatorjev in označevalca korpusa tudi lahko opravljajo vlogo diskurznega označevalca, pa v referenčni raziskavi (Verdonik et al., 2007a) niso bili obravnavani. Kateri od teh izrazov so bili označeni v validacijskih korpusih in kolikokrat, prikazuje tabela 6. Podatki v stolpcu K so za validirana korpusa, v stolpcih V1 za validacijski korpus prvega validatorja in stolpcih V2 za validacijski korpus drugega validatorja.

DO	Turdis-2			BNSIint		
Medmeti	**K**	**V1**	**V2**	**K**	**V1**	**V2**
hm	1	1	1	1	1	1
ma	1	1	1	1	1	0
a	1	0	2	0	0	0
he	0	0	0	1	1	1
vvv	0	0	0	0	0	1
evo	1	0	0	0	0	0
da	0	0	0	2	0	0
SKUPAJ	**4**	**2**	**4**	**5**	**3**	**3**
Ostalo						
in	0	0	0	3	0	26
torej	0	0	0	0	6	12
pa	0	0	12	0	0	0
namreč	0	0	0	0	0	7
pač	1	0	3	0	0	4
pol	0	4	3	0	0	0
pol pa	0	1	0	0	0	0
potem	0	3	0	0	0	0
seveda	0	0	0	0	1	5
saj	0	2	1	0	0	0
ampak	0	0	1	0	0	1
odlično	0	0	1	0	0	0
važi	0	1	1	0	0	0

v bistvu	0	1	1	0	0	0
tako	0	1	0	0	0	0
vem	0	1	0	0	0	0
da	0	0	0	0	0	2
naj	0	0	0	0	0	1
skratka	0	0	0	0	0	1
bom rekel	0	0	0	0	0	1
moram reči	0	0	0	0	0	1
SKUPAJ	**1**	**14**	**23**	**3**	**7**	**61**

Tabela 6: Novi izrazi v vlogi diskurznega označevalca.

Iz tabele 6 vidimo, da niti med obema validatorjema ni skupnega mnenja, kateri "novi" izrazi ustrezajo definiciji diskurznih označevalcev in kako pogosto so v tej vlogi. Zlasti velike razlike so pri izrazih *in, torej, pa, namreč* ipd., katerih konektorska vloga je sicer že bila prepoznana v domači (npr. Gorjanc, 1998) literaturi, za tuje tem sorodne izraze (npr. v angl. *and, so*) pa je prepoznana tudi pragmatična vloga (Schiffrin, 1987). Razlike so nedvomno v veliki meri posledica tega, da validatorja nista imela na voljo vira, ki bi podrobneje obravnaval pragmatične vloge teh izrazov, brez dvoma pa lahko sklenemo, da kaže razširiti raziskavo diskurznih označevalcev na nekatere priredne veznike (*in, torej, pa, namreč* ...), členke (*pač, seveda* ...) in prislove (*pol, potem* ...). Prav tako so v tej vlogi očitno nekateri, sicer redko rabljeni medmeti (v našem primeru *hm, ma*).

5. Zaključek

V prispevku smo predstavili validacijo označevanja diskurznih označevalcev v korpusih Turdis-2 in BNSIint ter skušali oceniti, do kolikšne mere so rezultati korpusne analize diskurznih označevalcev, ki smo jih uporabljali v jezikoslovnih raziskavah, nevtralni, ter preliminarno oceniti, ali je shema za označevanje diskurznih označevalcev, predstavljena v (Verdonik et al., 2007a), dovolj natančna oziroma v katerih segmentih jo je treba dopolniti.

Ugotovili smo, da je v korpusu Turdis-2 večja variabilnost pri označevanju diskurznega označevalca *zdaj* ter v korpusu BNSIint pri označevanju *ne?*. Za ostale diskurzne označevalce ugotavljamo, da so večinoma homogeno označeni.

Pri nadaljnjem razvoju označevanja diskurznih označevalcev v jezikovnih virih bi tako kazalo dodatno pozornost nameniti predvsem *zdaj* in *ne?*, pa tudi diskurznima označevalcema *ja* in *glejte*, pri katerih zaznamo variabilnost označevanja do 5 %. Pri vseh teh bi bilo tudi smiselno dopolniti validacijsko analizo s primerjanjem položajev, v katerih so bili označeni kot diskurzni označevalci.

Največ pozornosti pa je treba posvetiti izrazom, ki v shemi, na kateri sta temeljila označevanje in validacija (Verdonik et al., 2007a), niso eksplicitno obravnavani. Kot kažejo rezultati validacije, so to predvsem nekateri (priredni) vezniki, pa tudi členki, prislovi in medmeti. V tej smeri bi bile potrebne dodatne natančne raziskave pragmatičnih vlog teh izrazov, na katerih bi temeljila nadaljnja nadgradnja sheme za označevanje diskurznih označevalcev.

6. Literatura

D. Blakemore. 2002. *Relevance and Linguistic Meaning: The Semantics and Pragmatics of Discourse Markers*. Cambridge: Cambridge University Press.

D. K. Byron, P. A. Heeman. 1997. Discourse marker use in task-oriented spoken dialog. *5th European Conference on Speech Communication and Technology (Eurospeech)*, Rodos, Grčija.

L. Carlson, D. Marcu, in M. E. Okurowski. 2003. *Current Directions in Discourse and Dialogue*, poglavje Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. Kluwer Academic Publishers.

J. E. Fox Tree. 2006. Placing *like* in telling stories. *Discourse Studies* 8(6): 723-743.

B. Fraser. 1996. Pragmatic markers. *Pragmatics,* 6/2, 167-190.

B. Fraser. 1999. What are discourse markers? *Journal of Pragmatics* 31: 931-952.

V. Gorjanc. 1998. Konektorji v slovničnem opisu znanstvenega besedila. *Slavistična revija* 46/4. 367–388.

P. Heeman, J. Allen. 1999. Speech repairs, intonational phrases and discourse markers: modeling speakers' utterances in spoken dialog. *Computational Linguistics,* 25(4).

E. Miltsakaki, R. Prasad, A. Joshi in B. Webber. 2002. The Penn Discourse Treebank. *Language Resources and Evaluation Conference'04*, Lizbona, Portugalska.

R. Mitkov, R. Evans, C. Orasan, C. Barbu, L. Jones, in V. Sotirova. 2000. Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies. *Proc. of the Discourse Anaphora and Anaphora Resolution Colloquium (DAARC 2000)*, Lancaster, Vel. Britanija.

C. Müller, St. Rapp, in M. Strube. 2002. Applying co-training to reference resolution. *Proc. of the Annual Meeting of the Association for Computational Linguistics*, Philadelpha, ZDA.

A. Pisanski Peterlin. 2005. Text-organising metatext in research articles: An English-Slovene contrastive analysis. *English for Specific Purposes* 25: 307–319.

G. Redeker. 1990. Ideational and pragmatic markers of discourse structure. *Journal of Pragmatics* 14: 367-381.

D. Schiffrin. 1987. *Discourse Markers.* Cambridge: Cambridge University Press.

M. Schlamberger Brezar. 2007. Vloga povezovalcev v govorjenem diskurzu. *Jezik in slovstvo* 52(3-4): 21-32.

L. Schourup. 1999. Discourse markers. *Lingua* 107: 227-265.

M. Smolej. 2004. Členki kot besedilni povezovalci. *Jezik in slovstvo* 49(5): 45-57.

D. Verdonik, M. Rojc. 2006. Are you ready for a call? – Spontaneous conversations in tourism for speech-to-speech translation systems. *Proc. of 5th LREC*, Genova, Italija.

D. Verdonik, M. Rojc in M. Stabej. 2007a. Annotating discourse markers in spontaneous speech corpora on an example for the Slovenian. *Language Resources and Evaluation,* 41: 147–180.

D. Verdonik, A. Žgank in A. Pisanski Peterlin. 2007b. Diskurzni označevalci v dveh pogovornih žanrih. *Jezik in slovstvo,* 52/6, 19–32.

D. Verdonik, A. Žgank in A. Pisanski Peterlin. V tisku. The impact of context on discourse marker use in two conversational genres. *Discourse Studies.*

A. Žgank, T. Rotovnik, M. Sepesy Maučec, D. Verdonik, J. Kitak, D. Vlaj, V. Hozjan, Z. Kačič, B. Horvat. 2004. Acquisition and annotation of Slovenian Broadcast News database. *Proc. of 4th LREC*, Lizbona, Portugalska.