COGNITIVE STUDIES | ÉTUDES COGNITIVES, 14: 121–129
SOW Publishing House, Warsaw 2014

DOI: 10.11649/cs.2014.011

JOANNA TRZEBIŃSKA[1,A] & **JAKUB BARTOSZEWICZ**[1,2]
[1]Adam Mickiewicz University, Poznan, Poland
[2]Poznan University of Technology, Poznan, Poland
[A]joanna.trzebinska@amu.edu.pl former name: Joanna Szwabe

Multi-level Annotation of the Specialized Corpus of Dialogs of Disabled Polish Speakers

Abstract

While Polish language is relatively well represented in general purpose corpora such as National Polish Language Corpus still there are groups of speakers that are underrepresented in reference corpora. One of such sub-groups is the disabled people community. On the other hand there is a growing need for understanding how disability influences social and cognitive abilities, language in particular. In this paper, we present a specialized Corpus of Dialogs of Disabled Speakers. The process of compiling, transcription and annotation of pragmatic, semantic and morphosyntactic features will be described, as well as Corpus applications will be discussed.
Keywords: speech corpus, pragmatic annotation, semantic annotation, disability.

Introduction

The Corpus of Dialogs of Disabled Speakers (in Polish: Korpus Mowy Osób Niepełnosprawnych) has been designed and compiled in the course of the "Evaluation of the Situation, Needs and Competence of Polish Disabled People on a Sample of 10000 Individuals with Impairments" (in Polish: "*Ogólnopolskie badanie sytuacji, potrzeb i możliwości osób niepełnosprawnych na próbie 10000 ON*") supported by the National Fund for Rehabilitation of Disabled People (PFRON) in cooperation with the University of Social Sciences and Humanities in Warsaw and the European Social Fund. The project has been supervised by Anna Brzezińska, Adam Mickiewicz University in Poznan. The Corpus of Dialogs of Disabled Speakers (CDDS) has been compiled and annotated within "The Corpus Analysis of Disabled Speakers Utterances" module conducted by a team led by Joanna Trzebińska.

The literature on the mutual impact of language and disability is scarce and devoted mostly to mentally disabled (Happé, 1993; Langdon et al., 2002; Woźniak, 2000). However, the language of people with physical impairments has been

studied from different perspectives and with various methodologies. Some experimental research has demonstrated no differences in specific metaphor use by blind and healthy individuals of various age (Antović et. al., 2013; Minervino et al., 2009). On the other hand, neural activity patterns of adult patients suffering from traumatic brain injury have been shown to differ significantly from the control group during metaphor processing (Yang et al., 2010). Social impact of the notion of disability have also been studied, both by surveying the person-first language preferences of the concerned groups (Bickford,, 2004) and discussion of particular metaphors associated with disability (Vidali 2010). What is more, there has been some effort to design disability specific tools for evaluation of language development in case of children with motor and visual disabilities (Hennesey, 2011) and for providing a tailored Cognitive-Behavioral Therapy for patients suffering from medically unexplained symptoms (Sumathipala, 2013).

So far, there were no broad-scale corpus studies of the language of people with physical disabilities. As the social awareness of the problem of disability has been raising in the last two decades and the disability community itself has become interested in the research of their language and its socio-cognitive impact there is a growing need for studies of this kind. In this study, people with psychical or intellectual disabilities constituted a minor fraction of the sample, so it was possible to concentrate on the language of the physically disabled, providing a better insight in their language and cognition.

Structure of the CDDS

Data characteristics and format

20 group interviews featuring 113 subjects have been transcribed and annotated. The corpus consists of 402,146 units, including 225,299 words of raw text, with nearly 100 tags providing metadata concerning various features of both verbal and non-verbal communication. Detailed data characteristics is given in Table 1.

Table 1 Data characteristics

Word count	
without tags	225299
with tags	402146
Utterance count	18518
Sentence count	47356
Type-token ratio for words	0.0803
Type-token ratio for tags	0.0058
Sentence count to word count ratio	0.1347
Utterance count to sentence count ratio	0.6094

The corpus data consists of natural speech samples transcribed and stored in several formats including xml. Extensible markups and stylesheets allow easy tagging, filtering, transformation and retrieval of the raw text. An additional set of

scripts provides tools for automatic format conversion and correct concordancer input preparation.

Figure 1 Text sample in the xml tree form

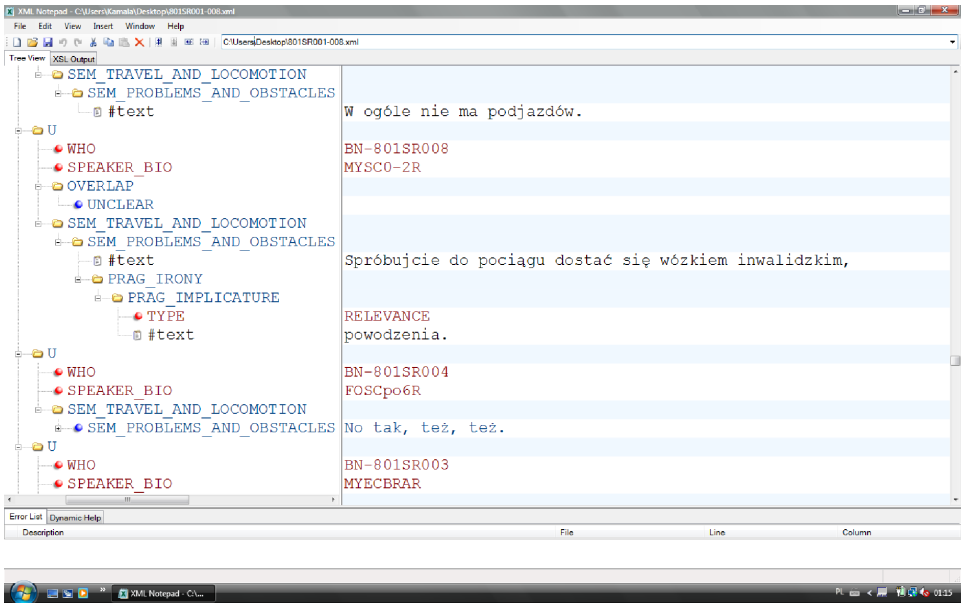


Figure 2 Raw text sample retrieved using xsl

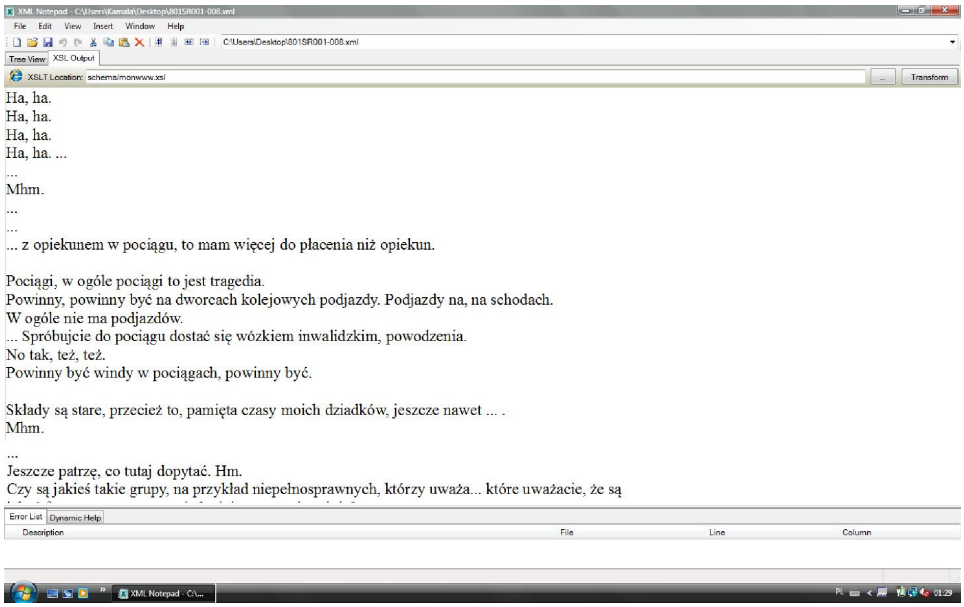


Figure 3 Text sample transformed for concordancing

```
<BN-801SR008> <SPEAKER_BIO=MYSC0-2R> <OVERLAP> <UNCLEAR/> </OVERLAP> <SEM_TRAVEL_AND_LOCOMOTION>
<SEM_PROBLEMS_AND_OBSTACLES> Spróbujcie do pociągu dostać się wózkiem inwalidzkim, <PRAG_IRONY>
<PRAG_IMPLICATURE=RELEVANCE> powodzenia. </PRAG_IMPLICATURE> </PRAG_IRONY>
</SEM_PROBLEMS_AND_OBSTACLES> </SEM_TRAVEL_AND_LOCOMOTION> </BN-801SR008>
<BN-801SR004> <SPEAKER_BIO=FOSCpo6R> <SEM_TRAVEL_AND_LOCOMOTION> <SEM_PROBLEMS_AND_OBSTACLES> No
tak, też, też. </SEM_PROBLEMS_AND_OBSTACLES> </SEM_TRAVEL_AND_LOCOMOTION> </BN-801SR004>
```

Annotation method

Corpora differ in both data types and tagging methods (Bougraev & Pustejovsky, 1996; Garside, Leech, McEnery, 1997; Lewandowska-Tomaszczyk, 2005; McEnery & Hardie, 2011). Corpus of Dialogs of Disabled Speakers is also annotated by its own system (Szwabe, 2009b). CDDS has been tagged with semantic, pragmatic and extra-linguistic metadata. Structural metadata, like partition of the corpus into dialogs, and dialogs into utterances, pauses and events, or group type (healthy, disabled, mixed), recording date and place, interview ID and number of interlocutors have also been tagged.

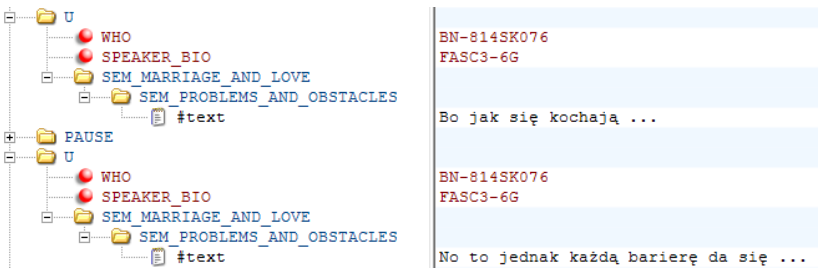
Every utterance is marked by an individual attributes of the speakers (WHO) and their type (SPEAKER_BIO). The first corresponds to the specific interview the speaker participated in, her or his role in the dialog (moderator — PR, healthy — BK, disabled — BN) and a personal ID. The latter consists of data concerning speaker's sex, place of residence (city, town, village), education, age group (18–30, 31–45, 46–73), disability type and onset. Disability acquisition time has been tagged using following intervals:

- Age $0-2$ — linguistic competence acquisition stage
- Age $3-6$ — pragmatic competence acquisition stage
- Age 6+ — mature form

This allows tracking utterances spoken by a specified type of speakers as well as analysis of effects demographic factors and case history may have on language acquisition and use. Subjects remained anonymous.

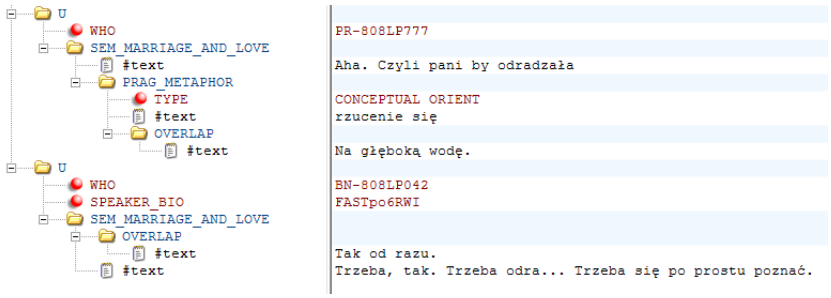
The semantic tagset consists of 40 tags coding semantic fields. They may be embedded if the semantic fields intersect, providing accurate data describing the topic of each utterance and semantic field co-occurrence tracking.

Figure 4 Semantic annotation sample



The pragmatic tagset consists of 9 tags allowing marking of pragmatic features like implicatures, metaphors, analogies, humor, irony, and indirect speech acts. Two perspectives have been applied to metaphor coding: conceptual metaphor theory (Lakoff & Johnson, 1980) and in parallel post-Gricean inferential model of communication (Grice, 1989). Four types of implicatures have been coded corresponding to Grice's Maxims of Quality, Quantity, Manner and Relevance.

Figure 5 Pragmatic annotation sample



The extra-linguistic tagset consists of 27 tags coding paralinguistic features (laughter, sigh etc.), utterance overlaps, pauses in the dialog, moments of silence after response-demanding acts of speech, syllabication and spelling, external events influencing the dialog's course, communicative gestures and facial expressions, singing, foreign words and unclear speech fragments, as well as structural[1] and biographical metadata.

In addition original forms such as typical errors, untypical errors, neologisms and region-specific forms have been marked and supplemented with their standard counterparts.

Furthermore, CDDS has been tagged morphosyntactically using Morfeusz automatic analyzer and TaKiPi tagger (Woliński, 2004). However, in numerous cases the resulting annotation included mutually exclusive tags, impairing further study requiring concordancer use. A short script choosing the most probable tag has been used to prepare the text for statistical analysis.

Other tools

Trancription has been performed using text editors supporting xml tagging and a previously prepared CDDS xsl template, allowing raw corpus text preview. Additionally, an xml schema, assisting transcription errors finding, and a file template providing consistent transcript format have been created. As word frequency lists and concordances are the elementary tools of corpus analysis, a script transforming the corpus text into a form accepted by most text analysis software (e.g. concordancers, frequencers) has been written.

[1]The discourse structure tags are compatible with TEI markup standard. According to TEI guidelines, TEI tagsets are designed to be extensible and therefore they may be combined, modified and redefined (TEI Consortium 2013).

Speakers characteristics

Utterances of 116 speakers, including 34 healthy speakers, 79 speakers with disabilities and 3 moderators are present in the corpus. Subjects were aged 18–73; having various levels of education (mostly secondary and higher education among the speakers with disabilities), coming from various parts of Poland, mainly middle-sized cities. The sex ratio has been close to 1, with a little overall male domination and female domination within the healthy group.

Speakers with disabilities have been divided into sub-groups according to their disability type: motor, visual, speech-auditory-vocal, psychic, mental. The groups are not disjoint as some of the subjects suffering from multiple conditions, fall into more than one group. The motor disability group has been the largest, the second being the visual disability group. What is more, motion and visual perception are the most important experiential bases used in figurative language by healthy language users. Thus, there is a good reason and opportunity to study figurative language use in those groups. Table 2 shows the speakers characteristics.

The corpus in the video version has been compiled using group interviews recorded in 2009. Subject recruitment rules are described by Iwański in a research report (Iwański, 2009, Iwański & Owczarek, 2010) containing more specific data concerning the subjects, and will not be discussed here. However, it should be mentioned that individuals whose disability "could prevent them from manifesting elementary communicative competence — like listening, talking, maintaining eye contact with an interlocutor" have been excluded from the study (Iwański, 2009).

Future development

It should be noted that pragmatic annotation is still rare in the corpus linguistics field. The corpus includes pragmatic annotation enabling study of language use from the cognitive pragmatics perspective (Szwabe, 2009a).

As the CDDS is semantically annotated by human taggers it may be used as a test-bed or a training set for machine learning algorithms in the process of semantic tagging automation for natural language corpora. The Corpus has already been used in an application of this kind: it has been shown that Latent Semantic Analysis based on Randomized Singular Value Decomposition may reduce the human effort necessary to semantically annotate a speech corpus in per-sentence tagging scenario. As a result of the study an automatic semantic tagger named Semancor has been designed, implemented and tested on samples of transcribed Polish speech derived from the CDDS itself (Prus-Zajączkowski et al. in prep.). Semancor has been successfully used for the Polish Child Speech Corpus annotation.

The primary use of the Corpus was a series of analyses of disabled people speech, conducted by Joanna Trzebińska in the course of the "Evaluation of the Situation, Needs and Competence of Polish Disabled People on a Sample of 10000 Individuals with Impairments" project. As the general results of the study show the differences between the disabled speakers and controls are found rather in communicative style than in linguistic competence, the corpus may be viewed as a supplementary reference corpus of Polish (Szwabe, 2009a).

Table 2 Speakers characteristics

								BN	**B**K	Total**
disability type		G*	I*	*P*	R*	U*	W*	no data		
number of subjects		13	25	10	63	3	27	2	34	177
sex	M	6	15	6	40	3	13	2	12	97
	F	4	10	4	23	0	13	0	22	76
age	Y	4	8	2	25	2	10	0	16	67
	A	3	6	5	18	1	8	1	6	48
	O	3	11	3	20	0	8	1	12	58
education	E	1	2	1	6	1	2	0	1	14
	S	7	17	6	38	2	17	0	14	101
	H	1	6	3	18	0	6	0	17	51
	no data	1			1		1	2	2	7
place of residence	T	3	12	4	25	2	15	0	6	67
	C	5	11	6	34	1	9	0	23	89
	no data	2	2		4		2	2	5	17
disability acquisition time	0 to 2	5	7	3	23	1	15	0	N/A	54
	3 to 6	1	0	0	2	0	0	0	N/A	3
	6+	3	15	6	33	2	9	0	N/A	68
	no data	1	3	1	5		2	2	N/A	14

*coincident with other disorders

**sum is greater than the total number of subjects because many of them suffer from more than one condition

BN — subjects with disabilities
BK — healthy subjects
G — speech-auditory-vocal
P — psychic
R — motor
U — intellectual
W — visual
I — other
M — male
F — female
Y — young adults
A — middle-age
O — mature
E — elementary education
S — secondary education
H — higher education
T — towns, villages
C — cities with 24,000+ inhabitants

References

Antović, M., Bennett, A. & Turner M. (2013). Running in circles or moving along lines: Conceptualization of musical elements in sighted and blind children. *Musicae Scientiae*, *17*(2), 229–245. doi: 10.1177/1029864913481470.

Bickford, J. O. (2004). Preferences of Individuals With Visual Impairments for the Use of Person-First Language. *RE:View*, *36*(3), 120–126. doi: 10.3200/REVU.36.3.120-126.

Bougraev, B. & Pustejovsky, J. (Eds.). (1996). *Corpus Processing for Lexical Acquisition.* MIT Press.

Garside, Leech, McEnery (Eds.). (1997). *Corpus Annotation.* London: Addison Wesley Longman.

Grice, H. P. (1989). *Studies in the Way of Words.* Cambridge, Mass., London: Harvard University Press.

Happé, F. (1993). Communicative Competence and Theory of Mind in Autism: a Test of Relevance Theory. *Cognition*, *48*(2), 101–119. doi: 10.1016/0010-0277(93)90026-R.

Hennesey, S. (2010). Assessing Early Language Development in Children with Vision Disability and Motor Disability. *International Journal of Disability, Development and Education*, *58*(2), 169–187. doi: 10.1080/1034912X.2011.570506.

Iwański, J. (2009). *Autodiagnoza grupowa (fokus) sytuacji życiowej i zawodowej osób z ograniczeniami sprawności Raport z badań w module nr 8 w ramach projektu pt.: Ogólnopolskie badanie sytuacji, potrzeb i możliwości osób niepełnosprawnych.* Warszawa: Szkoła Wyższa Psychologii Społecznej (unpublished manuscript).

Iwański, J., Owczarek, D. (2010). *Potrzeba bycia rozumianym. Komunikacja społeczna i funkcjonowanie w grupie osób z ograniczeniami sprawności.* Warszawa: Wydawnictwo Naukowe Scholar.

Lakoff G. & Johnson, M. (1980). *Metaphors we live by.* Chicago: University of Chicago Press.

Langdon, R., Coltheart, M., Ward, P., & Catts, S. (2002). Disturbed Communication in Schizophrenia: the Role of Poor Pragmatics and Poor Mind-reading. *Psychological Medicine*, *32*(07), 1273–1284. doi: 10.1017/S0033291702006396.

Lewandowska-Tomaszczyk B. (Ed.). (2005) *Podstawy językoznawstwa korpusowego.* Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice.* Cambridge: Cambridge University Press.

Minervino R. A., Martín A., & Trench M. (2009). Congenitally Blind do not Comprehend Better I Grasp the Idea than I See the Idea: A Challenge to the Use of Sensorymotor Conceptual Metaphors in the Comprehension of Metaphorical Expressions. In *Cognitive Science Conference Proceedings* (pp. 3004–3009).

Prus-Zajączkowski, B., Szwabe, J., & Szwabe, A. (in preparation) Automatic semantic annotation for natural language corpora — per-sentence tagging scenario.

Sumathipala, A. (2013). Development of metaphors to explain cognitive behavioural principles for patients with medically unexplained symptoms in Sri Lanka. *International Journal of Social Psychiatry*, *60*(2), 117–124. doi: 10.1177/0020764012467897.

Szwabe, J. (2009a). *Korpus Mowy Osób Niepełnosprawnych (MON) — Projekt struktury anotacji. Produkt 1 w ramach zadania Korpusowa analiza wypowiedzi osób niepełnosprawnych w module badawczym nr 8 — w projekcie pt.: Ogólnopolskie badanie sytuacji, potrzeb i możliwości osób niepełnosprawnych.* Warszawa: Szkoła Wyższa Psychologii Społecznej (unpublished manuscript).

Szwabe, J. (2009b). *Korpus Mowy Osób Niepełnosprawnych (MON) — Tagset. Produkt 2 w ramach zadania Korpusowa analiza wypowiedzi osób niepełnosprawnych w module badawczym nr 8 — w projekcie pt.: Ogólnopolskie badanie sytuacji, potrzeb i możliwości osób niepełnosprawnych.* Warszawa: Szkoła Wyższa Psychologii Społecznej (unpublished manuscript).

TEI Consortium, eds. *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* TEI Consortium. Retrieved October 30, 2013, from http://www.tei-c.

org/Guidelines/P5/.

Vidali, A. (2010). Seeing What We Know: Disability and Theories of Metaphor. *Journal of Literary & Cultural Disability Studies*, *4*(1), 33–54. doi: 10.1353/jlc.0.0032.

Woliński, M. (2004). Morfeusz — a Practical Tool for the Morphological Analysis of Polish. In M. Kłopotek, S. Wierzchoń, & K. Trojanowski (Eds.), *Intelligent Information Processing and Web Mining, IIS:IIPWM'06 Proceedings* (pp. 503–512). Berlin Heidelberg: Springer.

Woźniak, Tomasz. (2000). *Zaburzenia języka w schizofrenii*. Lublin: Wydawnictwo UMCS.

Yang, F. G., Fuller J., Khodaparast N., Krawczyk D. C. (2010). Figurative language processing after traumatic brain injury in adults: A preliminary study. *Neuropsychologia*, *48*(7), 1923–1929. doi: 10.1016/j.neuropsychologia.2010.03.011.

Publisher: Institute of Slavic Studies PAS & University of Silesia in Katowice