Review

Long-term Preservation of Longitudinal Statistical Surveys in Psycholinguistic Research

Hrvoje Stančić
Department of Information and Communication Sciences,
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, Zagreb, Croatia
hstancic@ffzg.hr

Martina Poljičak Sušec
Croatian Bureau of Statistics
Ilica 3, Zagreb, Croatia
poljicakm@gmail.com

Anabela Lendić
Department of Linguistics,
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, Zagreb, Croatia
alendic@ffzg.hr

Summary

Psycholinguistics deals with different types of evidence and obtained data, including confidential information which needs to be protected from disclosure and other security threats. When it comes to speech-language pathologies, researchers in psycholinguistics are especially interested in aphasia. Aphasia is a loss of language ability as a consequence of brain damage, which may result from head injury or stroke. Research data has to be adequately stored, processed, protected and if possible, preserved for secondary use. Authors are proposing possible application of models and tools used in official statistics and concepts from the archival science that could contribute to solving the so far unresolved issues in the research on aphasia and its records management requirements in the context of long-term preservation, trust, and reuse.

Keywords: psycholinguistics, statistics, archival science, records management, trust in digital records, anonymization, pseudoanonymization

Introduction

Psycholinguistics can be defined as "the study of mental representations and processes involved in language use, including the production, comprehension and storage of spoken and written language" (Warren 2013, 4). The interdisciplinary nature of the field is reflected in the fact that psycholinguistics is influ-

enced by research (and methodologies) in psychology, neurology, and linguistics, as well as those in cognitive science, computer science, and philosophy (Erdeljac 2009; Warren 2013). Since "psycholinguistics tends to blend the theoretical and descriptive insights of linguistics with the experimental methodology and rigor of psychology" (Warren 2013, 6), psycholinguistics has to deal with different combinations of types of evidence and obtained data. When it comes to speech-language pathologies, researchers in psycholinguistics are especially interested in aphasia. According to Warren (2013, 236), aphasia is: "loss of language ability as consequences of brain damage, which may result from head injury or stroke". In the field of psycholinguistics, aphasia research necessarily begins with researchers' request for access to aphasic subjects taking part in clinical therapy[1], and the research proposal needs to include a standard informed consent form approved by an ethic committee. The two most widely used diagnostic instruments for aphasia assessment in English-speaking countries are Boston Diagnostic Aphasia Examination (2001) and Psycholinguistic Assessment of Language Processing in Aphasia (PALPA) (1992). Practical comparisons and normalizations cannot be carried out in aphasia research when it comes to different languages, and data obtained from aphasia research in other languages can only serve as a (more or less "useful") general guideline when it comes to Croatian. There have been no large-scale studies of aphasia in Croatian so far. It is very important to note here that by agreeing to participate in a specific research, the participants also agree to share some of the personal information with the researchers. This paper analyses issues relevant in the context of the development of longitudinal studies on aphasia, of enabling secondary use of research data, and of long-term preservation of the collected data as authentic, reliable, and usable records with their integrity intact.

Psycholinguistics – legal and ethical aspect of protecting sensitive data

Protection of privacy, protection of personal data, and protection of private and family life is a basic human right[2] and is included in the European data protection laws[3]. Medical data are especially sensitive as their disclosure can harm persons by contributing to a social stigma and possible restriction or limitation of persons' rights. The Act on Personal Data Protection[4] regulates the protection of personal data regarding natural persons and the supervision of collecting,

[1] The most prominent clinical institution engaged in aphasia treatment in Croatia is The SUVAG Polyclinic, Section for Therapy of Speech Disorders – Speech Pathology, http://www.suvag.hr/en/, http://www.suvag.hr/en/sluzba-za-govorne-poremecaje-logopedija/.

[2] EU Parliament, The Council and the Commission, 2010.

[3] European Union Agency for Fundamental Rights, 2014.

[4] Legislation Committee of the Croatian Parliament, 2012.

processing, and using personal data in the Republic of Croatia. Concrete actions to prevent data disclosure are taken in the institutions handling data collection and other processes involving usage of sensitive data[5]. The Official Statistics Act[6] regulates official statistical system of the Republic of Croatia[7] and its principles, as well as confidentiality and data protection among other topics.

In aphasia research information such as gender, age, handedness, place of birth, level of education, and knowledge of other languages has to be included alongside other sources of data[8], rather than being separated from them. However, while some of the personal information such as those noted above can be provided by the participant himself, some of the information about the patient, for example his medical diagnosis (type of aphasia), needs to be provided by the clinical institution carrying out the therapy. Over time, the patient's medical record accumulates significant personal information including identification, history of medical diagnosis, the received treatment, medication history, psychological profile, and physicians' subjective assessments of personality and mental state, among others[9].

Abiding to the legal regulations and protection of privacy, researchers from different fields participating in a specific research should be able to access collected data on isolated aspects based on their research agendas.

Data collection – classification in clinical research

Classification in the area of medicine can be defined as a system of categories with criteria sorting diseases and conditions (or disorders, interventions in health, their costs, etc.) in discrete groups. A statistical classification is a classification having a set of discrete categories, which may be assigned to a specific variable registered in a statistical survey or in an administrative file, and used in the production and presentation of statistics[10]. Psycholinguistics should be able to consistently collect, analyse, compare, and interpret different language-related phenomena needed for statistical research processes. The methods of statistical classification could be applied in order to facilitate that process.The World Health Organization (WHO) is the main authority for classification systems in medicine and is responsible for the family of classifications used for statistical research, such as the International Statistical Classification of Diseases and Related Health Problems (ICD), International Classification of Func-

[5] Poljičak & Stančić, 2014.

[6] Official Gazette, No. 12, 2013.

[7] Hrvatski statistički sustav (HSS), http://www.dzs.hr/.

[8] For example, audio and video recordings of the participants performing language-related tasks.

[9] Mercury, 2004.

[10] Hoffmann & Chamie, 1999.

tioning, Disability and Health (ICF), and International Classification of Health Interventions (ICHI). ICD is widely used for mortality and morbidity statistics. Its 10th revision has been used by WHO Member States since 1994. The International Classification for Health Accounts (ICHA) forms a basis for the System of Health Accounts (SHA), an internationally accepted tool developed jointly by OECD, The European Commission, and WHO, and used for describing, summarizing, and analyzing expenditure on health and its financing. The Eurostat's "European shortlist" of 86 causes of death is based on ICD10. The International Shortlist for Hospital Morbidity Tabulation (2005) is likewise based on ICD-10 and is used by Eurostat, WHO, OECD and NOMESCO to collect and present data on hospital discharges[11]. In the course of a statistical survey there can appear a need to include other sorts of classifications, such as Nomenclature of Territorial Units for Statistics (NUTS[12]) to sort and analyze data by territory units. The Neuchâtel Terminology Model (NTM[13]) provides the conceptual framework for the development of a classification database. It defines the key concepts relevant for structuring classification metadata. NTM belongs to the semantic and conceptual sphere of metadata, and it does not include object types and attributes related solely to the technical aspects of a classification database.

Data should be classified according to the sensitivity levels and disclosure risks involved. Sensitivity is a measure of how freely the data can be handled. The sensitivity of a resource falls into two categories: restricted and unrestricted[14]. Data can be otherwise classified as confidential data, internal/private data, and public data[15]. Data (variables) can be classified as identifier, quasi-identifier, sensitive attributes, and non-sensitive attributes[16].

Data processing – preserving information and protecting privacy

Although it might seem contradictory, it is possible to both preserve collected information and protect patients' privacy, while at the same time making the results available to the researchers for further investigations. This can be achieved by preserving the raw data, or original datasets, and creating the processed datasets still retaining all characteristics of the original datasets, except for the pri-

[11] International classification, European Commission http://ec.europa.eu/health/indicators/international_classification/index_en.htm.

[12] Nomenclature of territorial units for statistics (NUTS), Eurostat, http://ec.europa.eu/eurostat/web/nuts/overview.

[13] Hustof, Born, Dunstan, & Mair, 2013.

[14] Sensitivity and Criticality of Data, 2010.

[15] Data Classification and Handling Policy, 2011.

[16] Fung, Wang, Chen, & Yu, 2010.

vacy-related variables. The first and the most basic step in maintaining privacy is to remove variables such as name, social security number, and home address. Removing obvious identifiers from the data is not always adequate to maintain privacy of an individual. Therefore, more rigorous procedures are required to achieve privacy. After removing obvious identifiers, some of the most basic methods for maintaining privacy include limitation of details, top/bottom coding, suppression, rounding and addition of noise[17]. After the data disclosure control methods have been applied, access to confidential data (identifiers and identifying variables) should be possible only if there is a reasonable need and authorization level required to access the variables. Access rights should be administered according to the sensitivity levels and authorization privileges for the data. The second step is the creation and implementation of specialized management system to handle data sensitivity levels and access rights.

Statistical metadata systems play a fundamental role in statistical organizations. A statistical metadata system (SMS[18]) is an important tool for ensuring the goals of the statistical information system are met. SMS should go beyond the function of support for production of official statistics to address other requirements. It should be a tool to facilitate efficient functioning and further development of the whole statistical information system. The possibility to assign access rights to some objects or variables of the database can be implemented on a metadata level to define and manage users and groups of users. Authorization rules can be assigned on specific items of metadata[19] and different levels of access can also be implemented. However, the ethical and legal constraints placed on data collectors processing medical records often limit secondary use of the material, even when anonymized, if prior legal consent has not been obtained[20]. Nevertheless, a system with functionalities similar to those of SMS could be used to manage sensitive health-related data.

In statistical surveys different specialists could be involved in statistical processes. Official statistics is constantly evolving in order to harmonize and standardize the area and even develop official statistics into a standardized industry for production and dissemination of official statistics' data. Usage of the statistical metadata system enables metadata-driven approach to architecture[21] and integration of applications[22]. However, sometimes integration is not possible, or it does not amount to positive outcomes. In these cases a balance be-

[17] Matthews & Harel, 2011.

[18] UNECE, 2009.

[19] Ryals.

[20] The Royal Statistical Society; the UK Data Archive, 2002.

[21] Rivera, Wall, & Glasson, 2013.

[22] Zeila, 2009.

tween integrated and tailored approach to some processes should be met[23]. These areas often include data editing and imputation in statistics. High-Level Group for the Modernisation of Statistical Production and Services (HLG)[24] created in 2010 by the Bureau of the Conference of European Statisticians has a mandate to reflect on and guide strategic developments in the ways in which official statistics are produced. In November 2012 HLG decided that developing Common Statistical Production Architecture (CSPA) was a key priority. Vale states that CSPA "builds on existing standards such as the Generic Statistical Business Process Model (GSBPM) and the Generic Statistical Information Model (GSIM) to create an agreed set of common principles and standards designed to promote greater interoperability within and between statistical organizations"[25]. It was developed to provide the "industry architecture" for official statistics. The Generic Statistical Information Model (GSIM[26]) is a conceptual model that provides a set of standardized, consistently-described information objects, which are the inputs and outputs in the design and production of statistics. Generic Statistical Business Process Model (GSBPM[27]) is a flexible tool to describe and define the set of business processes needed to produce official statistics. The value of the architecture is that it enables collaboration in developing and using services. One of the key enablers for the CSPA is Catalogues. It is envisaged that each statistical organization will have catalogues of processes, information objects, and statistical services. There will be a Global Artefact Catalogue which contains the shareable/reusable artefacts (i.e. processes, information objects and statistical services) from the statistical organizations. The Architecture Sprint proposed eight ways in which CSPA may be used[28].

Data preservation

As seen from the discussion so far, the official statistics has developed a sophisticated ecosystem of the models used by statistical organizations. Those models or parts of them (or even concepts) could be used in collecting, processing, and preserving health-related digital records. Further we will focus on the issues of the data storage and preservation. Data versioning is one of the concepts important in the context of preservation of authenticity of preserved records.

[23] Seljak, 2013.

[24] HLG, http://www1.unece.org/stat/platform/display/hlgbas.

[25] Vale, 2014.

[26] UNECE, 2013.

[27] Vale, Generic Statistical Business Process Modev v4.0, 2009.

[28] UNECE, 2013.

Versioning of statistical surveys in longitudinal studies should be possible using statistical metadata system in an information system. The collected data should be stored in the standardized, globally accepted file formats that enable long-term preservation and preservation of the authenticity and integrity of records, as well as assuring their accessibility. Retention policies for clinical records are defined by the legal framework of each country. Therefore, the minimum period of preservation should be set accordingly. The legal regulations vary considerably, but some types of records could be destroyed after several years, while others should be kept for 70 or more years. Nevertheless, many health care organizations hold records longer than mandated, but over time much of the clinical data become difficult to access[29]. In the context of digital data, long-term preservation means a significant number of media migrations and file type conversions during the preservation period. However, the data should stay authentic, reliable, usable, and its integrity should stay preserved at all times.

Institutions producing statistical survey data can manage data preservation with their own resources – persons with adequate skills to handle data preservation, technological requirements, and data storage facilities. Another approach would be to store the data in a digital, cloud-based archive and consider options available in that case. A Data Management Plan should be produced for all research projects that are creating or capturing data[30].

The choice of metadata standard to be used for data documentation is crucial for ensuring interoperability with other institutions' systems. Prevailing standards for data documentation in social sciences are presented in Table 1 which shows some of the available metadata standards.

Table 1: Standards for depositing social science data

Disciplinary area	Metadata standard	Description
Social Sciences	Data Documentation Initiative (DDI)	A metadata specification for the social and behavioral sciences created by the Data Documentation Initiative. Used to document data through its lifecycle and to enhance dataset interoperability.
	Statistical Data and Metadata Exchange (SDMX)	A self-describing data format that provides both metadata and a method of data transmission. It is primarily used in "the world of official statistics", such as the EU, WHO, UNESCO, World Bank, and US Reserve Banks.

Source: Data Management Resources and Services, University of Idaho Library, http://www.lib.uidaho.edu/services/data/data-management/#standards.

[29] Corn, 2009.

[30] London School of Hygiene and Tropical Medicine, 2014.

Both, SDMX and DDI-L (i.e. DDI-Lifecycle), as presented in Figure 1, are dedicated to repurposing data in social sciences and enabling re-use of results for secondary analysis.

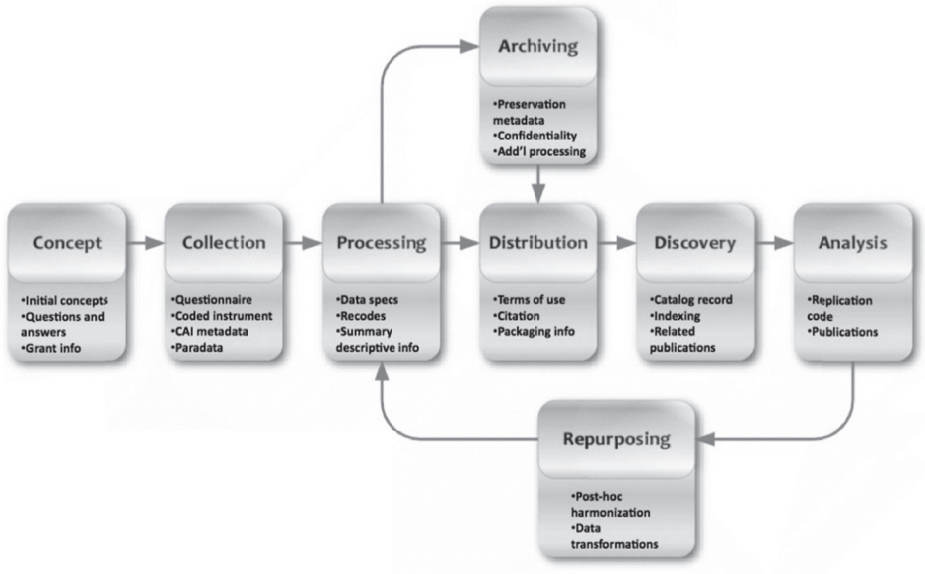

Figure 1. The DDI-Lifecycle Model fostering metadata reuse. DDI Alliance, http://www.ddialliance.org.

The ethical and legal constraints placed on data collectors processing medical records often limit secondary use of the material, even when anonymised, if prior legal consent has not been obtained[31]. The *Processing* phase of the DDI-L Model is where the data about patients is to be properly described. If this phase is done correctly, the anonymization and pseudoanonymization could be achieved. This phase is also the basis for further research made possible through the *Repurposing* phase. Although the two mentioned standards facilitate data and metadata reuse, the key issue is to enable longitudinal studies on aphasia and secondary use of research data. By applying the DDI-L Model to the records management in the aphasia research it could be possible to avoid privacy issues.

[31] The Royal Statistical Society; the UK Data Archive, 2002.

Conclusion

This paper aims to show interrelationships between the needs of psycholinguistic research and the available models and standards in official statistics. Solutions considered range from the ones for data collection, statistical survey processing, and records management. Official statistics' tools can be used to support interdisciplinary research with underlying processes correlating to official statistics' processes in statistical surveys. Furthermore, practices in documenting information objects can be useful as well, as they are based on conceptual frameworks and models – enabling reuse of artefacts. It is the generic nature of standards and models in official statistics and strong metadata orientation, as well as experience in metadata systems' implementation that is seen as valuable to consider for involvement in psycholinguistics in order to have strong methodological ground for data documentation, as well as for enforcement of sensitive data protection and risk management. When it comes to the area of personal data protection in psycholinguistic research, we believe that the records creators can utilize the appropriate methods for statistical disclosure control in planning, realizing, and using the aphasia records management system. A system built in the way in which it is possible to control the metadata, and in turn control the sensitive records, could be than freely used by researchers. By being able to access data and records with no restrictions, the researchers will be able to improve the level of knowledge about aphasia and the patients will be provided with a more successful treatment. Also, if realised as proposed, the records management system will gain trust by patients', their families, and researchers since it would implement all the necessary standards, methods, procedures, metadata schemes and control mechanisms already tested and trusted. Therefore, it can be concluded that the aphasia-related psycholinguistic field of research will become more productive if it incorporates knowledge and solutions from official statistics and modern archival science. We strongly believe that by combining well known and established methods of official statistics with long-term digital records preservation methods and applying them to the field of aphasia research, the reuse of collected patients' data, longitudinal studies on aphasia, secondary use of research data, and protection of patient' privacy rights could be seamlessly achieved.

References

London School of Hygene and Tropical Medicine. (2014.). *Research Data Management Policy.* Retrieved 2015., from http://researchonline.lshtm.ac.uk/612422/1/RDM-Policy-v10.pdf

Corn, M. (2009.). *Archiving the Phenome: Clinical Records Deserve Long-term Preservation.* Retrieved 2015., from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2605592/pdf/1.S1067502708001849.main.pdf

Erdeljac, V. (2009). *Mentalni leksikon: modeli i činjenice.* Zagreb: Ibis grafika.

EU Parliament, The Council and the Commission. (2010). *Charter of Fundamental Rights of the European Union.* Retrieved 2015., from http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.C_.2010.083.01.0389.01.ENG

European Union Agency for Fundamental Rights. (2014.). *Handbook on European data protection law.* Retrieved 2015., from http://www.echr.coe.int/Documents/Handbook_data_protection_ENG.pdf

Fung, B. C., Wang, K., Chen, R., & Yu, P. S. (2010.). *Privacy Preserving Data Publishing: A Survey of Recent Developments.* Retrieved from https://www.cs.sfu.ca/~wangk/pub/FWCY10csur.pdf

Goodglass, H., & Kaplan, E. (2001). *Boston diagnostic aphasia examination.* Philadelphia, PA: Lippincott, Williams & Wilkins.

Hoffmann, E., & Chamie, M. (1999.). *Standard Statistical Classifications: Basic Principles.* Retrieved from http://unstats.un.org/unsd/class/family/bestprac.pdf

Hustof, A. G., Born, A., Dunstan, T., & Mair, D. (2013.). *NEUCHATEL TERMINOLOGY MODEL: CLASSIFICATION DATABASE OBJECT TYPES AND THEIR ATTRIBUTES. REVISION 2013.* Retrieved 2015., from http://www.google.hr/url?sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=1&ved=0CBwQFjAA&url=http%3A%2F%2Fwww.dst.dk%2Fext%2F595717909%2F0%2Fukraine%2FENG_Statistical-Metadata-Working-Paper-Part-7--pdf&ei=DO6PVf7NK4nWU5j7toAC&usg=AFQjCNHcwJrueqMojwloAB0iRl3IGk1h

Kay, J., Lesser, R., & Max, C. (1992). *Psycholinguistic Assessment of Language Processing in Aphasia (PALPA).* Hove, England: Lawrence Erlbaum Associates.

Legislation Committee of the Croatian Parliament. (2012.). *The Act on Personal Data Protection.* Retrieved 2015., from http://narodne-novine.nn.hr/clanci/sluzbeni/2012_09_106_2300.html

Matthews, G. J., & Harel, O. (2011.). *Data Confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy.* Retrieved from https://projecteuclid.org/download/pdfview_1/euclid.ssu/1296828958

Mercury, R. (2004.). *The HIPAA-potamus inn Health Care Data Security, Communication of the ACM, vol . 47, no.7.* Retrieved 2015.

Michigan Technological University Information Technology Services and Security. (2011.). *Data Classification and Handling Policy.* Retrieved 2015., from https://security.mtu.edu/policies-procedures/DataClassificationAndHandlingPolicy.pdf

Official Gazette No. 12. (2013.). Zakon o službenoj statistici. *Narodne novine*.

Poljičak, M., & Stančić, H. (2014.). *Statistical Microdata: Production of Safe Datasets, Transparent Presentation of Contents and Advanced Services for Users through Metadata Authorisation System.* Retrieved 2015., from https://bib.irb.hr/datoteka/714198.Poljicak_ Stancic_ Statistical_Microdata.pdf

Rivera, A., Wall, S., & Glasson, M. (2013.). *Metadata Driven Business Process in the Australian Bureau of Statistics.* Retrieved 2015., from http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.40/2013/WP21.pdf

Ryals, M. (n.d.). *SAS Metadata, Authorization and Management Services - Working Together for You.* Retrieved 2015., from http://www.lexjansen.com/nesug/nesug03/et/s1059.pdf

Seljak, R. (2013). Integrated statistical systems and their flexibility - How to find the balance. *NTTS - Conferences on New Techniques and Technologies for Statistics*, (pp. 269-278). Brussels.

Seljak, R. (n.d.). *Integrated statistical systems and their flexibility - How to find the balance.* Retrieved 2015., from http://www.stat.si/dokument/5287

The Royal Statistical Society; the UK Data Archive. (2002). *Preserving and Sharing Statistical Material.* Retrieved 2015, from http://www.rss.org.uk/Images/PDF/publications/rss-reports-preserving-sharing-statistical-data-2002.pdf

The Royal Statistical Society; the UK Data Archive. (2002.). *Preserving and Sharing Statistical Material.* Retrieved 2015., from rss-reports-preserving-sharing-statistical-data-2002

UNECE. (2009.). *Common Metadata Framework Part A: Statistical Metadata in a Corporate Context.* Retrieved 2015., from http://www1.unece.org/stat/platform/display/metis/The+Common+Metadata+Framework

UNECE. (2013.). *Fostering Interoperability in Official Statistics: Common Statistical Production Architecture.* Retrieved 2015.

UNECE. (2013.). *Generic Statistical Information Model.* Retrieved 2015., from http://www1.unece.org/stat/platform/display/gsim/GSIM+Specification

UNECE. (2013.). *Generic Statistical Information Model (GSIM): Specification.* Retrieved 2015., from http://www1.unece.org/stat/platform/display/gsim/GSIM+Specification

University of South Florida. (2010, 6 11). *Sensitivity and Criticallity of Data. Securiy Standard.* Retrieved 2015, from http://www.usf.edu/it/documents/issp001-3-3-10.pdf

University of South Florida. (n.d.). *Sensitivity and Criticallity of Data, Securiy Standard.* Retrieved 2015., from http://www.usf.edu/it/documents/issp001-3-3-10.pdf

Vale, S. (2009). *Generic Statistical Business Process Modev v4.0.* Retrieved 7 4, 2015

Vale, S. (2009.). *Generic Statistical Business Process Modev v4.0.* Retrieved 2015.

Vale, S. (2014). The Common Statistical Production Architecture: An Important New Tool for Process Standardisation. *European Conference on Quality in Official Statistics (Q2014).* Vienna.

Vale, S. (n.d.). *The Common Statistical Production Architecture: An Important New Tool for Process Standardisation.*

Warren, P. (2013). *Introducing Psycholinguistics.* New York: Cambridge University Press.

Zeila, K. (2009.). *Metadata Driven Integrated Statistical Data Processing and Management System.* Retrieved 2015.