The Journal of Chinese Linguistics Vol.46, No.2 (June 2018): 227-268

LINKING BASIC LEXICON TO SHARED ONTOLOGY FOR ENDANGERED LANGUAGES: A LINKED DATA APPROACH TOWARD FORMOSAN LANGUAGES

Chu-Ren Huang *The Hong Kong Polytechnic University*
Shu-Kai Hsieh *National Taiwan University*
Laurent Prévot *Aix-Marseille Université & CNRS, France*
Pei-Yi Hsiao *National Tsing Hua University, Taiwan*
Henry Y. Chang *Academia Sinica, Taiwan*

ABSTRACT

This paper proposes an innovative approach to link basic lexicon (e.g. Swadesh list) to upper ontology as the foundation of OntoLex interface to address the challenge of building language resources for endangered languages in the linked data paradigm. A linked data approach to language resources requires existing, and preferably sizable, language resources. For endangered and other less-resourced languages, however, the scarcity of existing resources limits the possibilities and potential benefits of linking. The challenges are then, how can construction of language resources for endangered language continue to thrive in the linked data paradigm, and how can the linked data approach benefit language resources for endangered languages. Our proposal requires the bare minimum of available data and we show with examples from Formosan languages (Austronesian or aboriginal languages of Taiwan (Blust 2013, 20))[i]

Authors claim no conflict of interests to publish this paper in *Journal of Chinese Linguistics*.

Chu-Ren Huang (corresponding author)
[churen.huang@polyu.edu.hk]; https://orcid.org/0000-0002-8526-5520
Shu-Kai Hsieh [shukaihsieh@ntu.edu.tw]; https://orcid.org/0000-0001-9674-1249
Laurent Prévot [laurent.prevot@lpl-aix.fr]; https://orcid.org/0000-0002-2463-2382
Pei-Yi Hsiao [hpy0804@gmail.com]; https://orcid.org/0000-0003-2870-7158
Henry Y. Chang [henryylc@gate.sinica.edu.tw]; https://orcid.org/0000-0002-3734-6772

i. The term "Formosan languages" conventionally refers to the Austronesian languages, not to the Sinitic languages, spoken in Taiwan ("…it is customary to use 'Formosan' to refer to the aboriginal languages of Taiwan. I follow this practice, and use 'Formosa' as a geographical designation for the pre-modern period, …" Blust 2013: 20).

that 1) this approach is applicable to endangered languages, and that 2) in spite of the restrictions imposed by scarcity of resources, the linked linguistic data consisting of basic lexicon + upper ontology generate important new information. Comparing Swadesh lists from different languages allowed us to build a small shared ontology that reflects direct human experience, and can serve as the cross-lingual conceptual core. In addition, these micro-ontologized lexicons can be used as seeds for developing a fully-grown and more comprehensive documentation of linguistically motivated ontology for each language.

KEYWORDS
Endangered languages **L**inked Data **S**wadesh list **O**ntology **S**UMO **F**ormosan languages (Austronesian languages in Taiwan)

1. INTRODUCTION

Language resources have witnessed a substantial growth and emergent diversity in recent years. Modeling the heterogeneity and multitude of language resources in an interoperable way has gained much attention in the research community of Language Resources (e.g. Stede and Huang 2012, McGrae et al. 2015). Although earlier work, such as those spearheaded by the Open Language Archives Community (OLAC)[ii], Bird and Simon 2003) focused on the sharability of metadata and accessibility of the resources through a common repository, recent trends in the Linked Data Paradigm (Chiarcos et al. 2012) in the context of Semantic Web (Berners-Lee 2006, Buitelaar and Cimiano 2004) poses both new opportunities and new challenges. The linked data approach requires that the content of the resources being accessible and interpretable under shared ontology in addition to the accessibility of the data. This has become a promising approach for cultural heritage and language documentation as it allows rich representation and preservation of cultural knowledge (Hyvönen 2012); it also poses serious challenges for less-resourced, and especially endangered languages, as they face the most pressing difficulty

ii. Open Language Archives Community (OLAC) http://www.language-archives.org/

The Journal of Chinese Linguistics Vol.46, No.2 (June 2018): 227-268

ZHANG, Huarui, Chu-Ren Huang and Shiwen Yu. 2004. Distributional consistency: A general method for defining a core lexicon. Paper presented at The Fourth International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal. In the proceedings of the Conference, 1119-1122, https://www.academia.edu/2069216/Distributional_consistency_A_general_method_for_defining_a_core_lexicon, accessed Jan. 16, 2018.

濒危语言基本词库与上层知识本体的链接
—关联数据在台湾南岛语研究的应用

黄居仁 香港 理工大学
谢舒凯 国立 台湾大学
龙安培博 Aix-Marseille Université & CNRS, France
萧佩宜 国立 清华大学, 台湾
张永利 中央研究 语言研究所, 台湾

提要

关联数据(linked data)研究法的兴起对濒危语言的语言典藏造成了极大的挑战。本文在本体词库界面(OntoLex)的基础上提出链接基本词库（如斯瓦迪斯词表 (Swadesh list)）与上层知识本体的新进路，藉以验证关联数据方法在濒危语言语言典藏的可行性。关联数据是在网路语意化后构建语言资源最重要的手段。但是关联数据法成功的前提需要有现成的大量语料或语言资源可以链接。把这个研究法应用到濒危或其他资源匮乏语言，所有关联数据的优势都会因为缺乏可链接的现成资源而消失殆尽。 在关联数据范式主导网路研究与资源构建的环境下，濒危语言典藏面临了如何在资源匮乏的劣势中，连接产生新资源与新知识的严峻挑战。本文以台湾南岛语为对象，提出仅需要最少资源的资源链接方法，以证实 1) 关联数据法可以用于濒危语言，2) 即便是资源匮乏，基本词库与上层知识本体的链接可以产生新的文化知识。比较斯瓦迪斯词表在不同语言中的呈现，使研究者可进一步在上层共享知识本体的架构下 比较不同语言文化间的基本概念体系与生活经验差异。 这些核心知识本体更可以作为未来为这些语言的构建完整知识本体的基础。

关键词

濒危语言 **关**联数据 **斯**瓦迪斯词表 **知**识本体 **S**UMO **台**湾南岛语

The Journal of Chinese Linguistics Vol.46, No.2 (June 2018): 227-268