國立臺灣大學電機資訊學院資訊工程學系

博士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering & Computer Science

National Taiwan University

doctoral dissertation

台語文處理技術：以變調及詞性標記為例

Processing Techniques for Written Taiwanese -- Tone Sandhi and POS Tagging

楊允言

Iûnn Ún-giân

指導教授：高成炎博士 陳克健博士

Advisors: Gao Cheng-yan Ph.D. Chen Keh-jiann Ph.D.

中華民國 98 年 1 月

January, 2009

國立臺灣大學博士學位論文
口試委員會審定書

台語文處理技術：以變調及詞性標記為例
Processing Techniques for Written Taiwanese -- Tone Sandhi and POS Tagging

本論文係楊允言君（學號 D93922001）在國立臺灣大學資訊工程學系完成之博士學位論文，於民國 98 年 1 月 16 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

高成炎	陳志昌
（指導教授）	陳信希
王旭正	高照明
張俊盛	劉昭麟
李琳山	余明興

系主任 呂育道

Preface

Taiwanese is my mother tongue. I lived in Taipei for 30 years, since I was born. I spoke my mother tongue before I entered elementary school. I could also understand Mandarin, but could not speak it well. However, after entering elementary school, I gradually forgot my mother tongue. I was able to speak Mandarin fluently, and I could understand Taiwanese but found it difficult to express my thoughts in my mother tongue. In school, they told me that Taiwanese was only a dialect, and that we should not speak it in school. In fact, if we spoke Taiwanese, we could be punished. I accepted this argument until I was 20 years old.

At that time, I enrolled at National Taiwan University (NTU), during the period when Taiwan was still under martial law. A political event occurred on campus that gave me a big shock. I could not concentrate on my studies for several months. I was gradually moving out of the shadows, establishing my own independent thinking and values. I realized that cultural enlightenment is more important than political enlightenment, and that our mother tongue is a solid foundation for Taiwanese culture.

I began to relearn my mother tongue. Then, I realized that it would not be enough to just be able to understand and speak the language. The ability to read and write the Taiwanese language could contribute to its promotion.

I began to read Taiwanese material, which is difficult to find in bookstores.

I also wrote letters to my friends in Taiwanese, mixing various kinds of scripts, including Han characters, phonetic symbols, KK phonetics, and Kana. Afterwards, I borrowed a book from the NTU library entitled, *The speech structure and transcription method of Taiwan Hokkianese* '台灣福建話的語音結構及標音法,' which was written by Tē^{n}, Liông-úi and published in 1978 . It took me four days to learn POJ. I also founded the NTU Puppet Shows Association '台大掌中劇團' with my classmate Tiu^{n}, Lân-sėk '張蘭石' during my last year of university life. I strongly advocated the use of written Taiwanese to write the drama scripts, but finally failed due to a lack of time

Based on my computer science background, my study interest shifted to natural language processing during my graduate years, from 1991 to 1993. I processed Mandarin. At the same time, I was still engaged in the written Taiwanese movement. My friends and I founded the "Students Taiwanese Promotion Association," (STAPA) '學生台灣語文促進會,' published the "Taiwanese Student" magazine '台語學生' (4 trial issues and 22 issues in total), wrote articles in Taiwanese, *etc*.

I entered the Dahan Institute of Technology CSIE Department as a lecturer and began to do Taiwanese processing related research work in 1999. Due to the lack of Taiwanese language related resources, I tried to establish some online written Taiwanese tools, including the Online Taiwanese-Mandarin Dictionary (OTMD), Online Taiwanese Syllable Dictionary (OTSD), and Online Taiwanese Concordancer System (OTCS), with more than 9-million-syllable written Taiwanese materials, Taiwanese word frequency reports with both

Han-Romanization mixed script and Taiwanese Romanization script, *etc*.

In my opinion, the computational linguistics related academic society is a minority in Taiwan. It is difficult for a researcher to master both computer science and the language itself. Obviously, Taiwanese processing related research is less popular in this minority, although I cannot understand why most researchers abandon their mother tongues. I selected Taiwanese language processing as my major field for study, tried to create this “new” field of research, and have achieved some preliminary results. I had the computer science background, and have been engaged in written Taiwanese work for more than 20 years.

I love my mother tongue and want to promote its social status. This is a difficult but, I believe, an important task.

Acknowledgments

I started reading and writing Taiwanese in 1987 when I was an undergraduate student, began written Taiwanese processing related research in 1999, and re-entered NTU CSIE as an adult student in 2004. I have always believed that working in this field is a worthwhile endeavor.

I am very grateful to my advisor, Professor Cheng-yan Kao, who encouraged me to select this topic and offered me some related research opportunities. My appreciation also goes to research fellow Keh-jiann Chen, my co-advisor, who generously provided me with many Chinese language processing resources. Many thanks to my committee members, Professor Jason S. Chang, Hsi-Jian Lee, Hsin-hsi Chen, Chao-Lin Liu, Ming-shing Yu, Khîn-huānn Lí, and Zhao-ming Gao, for their many constructive suggestions.

Thanks to my academic partner Ha̍k-khiam Tiunn. Thanks also to Professor Yuh-dauh Lyu and Doctor Ho̍k-chû Tiunn for correcting my written English in this dissertation. Thanks to Professor Jū-hông Tiu^{n} for helping me to understand some linguistic terms that confused me.

OK, I will leave space for my mother tongue. The following content will be written in Taiwanese.

Iōng ka-kī ê bú-gú siá, siá liáu khah kín, thang khah kín pit-gia̍p ☺

Ùi 1993 nî gián-kiù-só· pi̍t-gia̍p liáu-āu khai-sí kóng. Hit-chūn tùi lâu tī ha̍k-su̍t-kài si̍t-chāi bô chhù-bī, kám-kah chòe Tâi-gú-bûn ūn-tōng khah si̍t-chāi. Góa seng kè-sio̍k ku tī Tiong-gián-ī^{n} Sû-khò· sió-cho· chia̍h thâu-lō·, ná tih chòe Tâi-bûn kang-chok-chiá hóng-tâm. 1994 nî góa khì ín Tâi-tāi tiān-sòan tiong-sim ê thâu-lō·, kám-siā Tân Bûn-chìn lāu-su, thia^{n}-kóng góa ùi 60 kúi ê èng-cheng-chiá hông kéng--tio̍h, lāu-su ū tàu-sa^{n}-kāng. Tī hiâ chòe ê khah tiōng-iàu ê tāi-chì, tō-sī siá tāi-ha̍k liân-khó hun-hoat thêng-sek, che sī Khó· Sùn-khim lāu-su hō· góa ê ki-hōe.

Góa mā khai-sí peh-soa^{n}, chham-ka Siā Tiông-têng teng-san hōe, koh ū chòe

kàn-pō͘, sėk-sāi chin chōe hó pêng-iú. Kám-siā Kì Chhun-heng ê kài-siāu, in-ūi peh-soa^{n}, góa sėk-sāi Giȯk-lêng, liáu-āu kiat-hun. Bi̍t-gȯat ê tē it chām, sī tè Siā Tiông-têng teng-san-hōe ê pêng-iú chòe-tīn peh Giȯk-san, chòe chi̍t kiā^{n} tiōng-iàu ê tāi-chì, chit hāng tāi-chì, it-ti̍t kàu 12 tang āu chiah ū mûi-thé pò-tō. [1] Góa koh tȧk lé-pái kò͘-tēng chi̍t àm tī Tâi-ôan ê Tiàm chòe gī-kang, thâu-ke A-sam-ko kah Biāu-lêng-chí tùi Tâi-ôan ê ài, sī góa chiok kèng-pōe ê.

In-ūi peh-soa^{n}, jú lâi jú kám-kah tòa tī Tâi-pak si̍t-chāi sī chiok kan-khó͘ ê tāi-chì, lâng chōe, chhia chōe, khong-kan ȯeh, khong-khì bái. Góa khai-sí chhē só͘-chāi, ū soa^{n} koh lī to͘-chhī khah hn̄g ê, ū Hoa-liân, Tâi-tang kah Lâm-tâu. 1996 nî, in-ūi chhē-tio̍h Hoa-liân ê thâu-lō͘, tī lāu-pē lāu-bú kah Giȯk-lêng lóng bô chi-chhî ê chêng-hêng hā, kā Tâi-tāi ê thâu-lō͘ sî-tiāu, cháu lâi Hoa-liân.

1996 nî, Hoa-liân iah-koh ū pėh-sek khióng-pò͘ ê khì-hun, góa bô kā lí phiàn!

Hó ka-chài chi̍t tang āu Giȯk-lêng tiàu lâi Hoa-liân, só͘-í seng-ȯah thang khah ún-tēng. Gún tī Hoa-liân bóe chhù, Tan-hông kah Ka-chhái mā sio liân-sòa chhut-sì.

Seng-ȯah ún-tēng bô piáu-sī thâu-lō͘ ún-tēng, góa tī Hoa-liân, sa^{n} tang ōa^{n} gō͘ ê thâu-lō͘. Góa iah-sī hoat-kak, nā-sī beh chòe Tâi-gú-bûn, tī hȧk-su̍t-kài iah-sī khah hó-sè. 1999 nî, góa ji̍p-lâi Tāi-hàn ki-su̍t hȧk-ī^{n}.

Kám-siā Hȧk-khiam hia^{n}, i chin khòa^{n}-tiōng góa, ū-sî-á hō͘ góa ki-hōe khì i ê pan ián-káng/siōng-khò, mā ū chiap kè-ėk, pun keng-hùi hō͘ góa tàu chòe bāng-chām.

Góa khai-sí kiàn-li̍p Tâi-gú-bûn gián-kiù ê ki-chhó͘, Tâi-hôa sû-tián (kám-siā Tē^{n} Liông-úi kàu-siū kah chin chōe pêng-iú), Tâi-gú jī-tián, gú-liāu-khò͘ (kám-siā Chùn-iȯk-hia^{n}, Pek-nî hia^{n}, Tek-hôa, Lē-soat, …), Siȯk-gú (kám-siā Siau Pêng-tī lāu-su), Koa-iâu (kám-siā Gô͘ Jîn-sek lāu-su), Chhiò-khe-tāi (kám-siā Iû^{n} Chín-jū hāu-tiú^{n}, Siau lāu-su, Pek-nî hia^{n}) téng-téng.

Góa siū^{n}-beh kái-tô͘, 1999 nî pat thȧk kè Tong-hôa Chȯk-kûn só͘ chāi-chit choan-pan chi̍t hȧk-kî, 2002 nî khì khó Sêng-tāi Tâi-bûn-hē tē it kài phok-sū-pan,

[1] http://www.libertytimes.com.tw/2007/new/feb/13/today-fo5.htm

kám-siā Lîm Sìn-kian bo̍k-su thê-kiong sok-sià hō͘ góa tòa, gún sī chho͘-chhù kì^{n}-bīn. Sêng-tāi bô khó tiâu, tû-liáu ka-kī bô châi-tiāu, thia^{n}-kong koh-khah chú-iàu ê gôan-in sī lāi-tóe chi̍t kho͘ chiok chheh chòe Tâi-gú ê. Kám-siā Lī Heng-chhiong kàu-siū, i khó-lêng kám-kah ti̍t-chiap kā góa kóng khó phok-sū-pan bô bāng tùi góa ê tá^{n}-kek siu^{n} tōa, só͘-í kiàn-gī góa ùi góa gôan-lâi ê choan-gia̍p chit-pêng lâi seng-téng khòa^{n}-māi. Kám-siā Tē^{n} Liông-úi kàu-siū, i hiòng góa iau kó, siá chi̍t phi^{n} lūn-bûn tâu khì IJSL, khó-sioh āu--lâi chit pún Special Issue on Taiwanese pēng bô chhut-pán, kám-siā Hông-giâu (Henry), i tap-èng chòe-tīn siá chit phi^{n} lūn-būn, i ê Eng-bûn iáu-siū chán.

In-ūi ū chit phi^{n} lūn-bûn, góa tō sin-chhéng seng-téng, kám-siā Kán Li̍p-hong lāu-su, i siông-sòe khòa^{n} góa seng-téng ê chu-liāu, hō͘ góa chin chōe kiàn-gī, i kā góa kóng, in-ūi chin-chōe lâng bô liáu-kái góa tih chòe ê mi̍h-kiā^{n}, só͘-í seng-téng ê chu-liāu ài lēng-gōa siá chi̍t hūn Proposal, kau-tài Tâi-gú-bûn chū-jiân gú-giân chhú-lí gián-kiù ê tiōng-iàu-sèng. Seng-téng sàng-sím ê sî-kan, tán chin kú chiah ū siau-sit, thia^{n}-kóng ū úi-ôan bôai sím, thòe tn̂g-khì Ha̍k-sím-hōe. Góa chin hó-ūn, sūn-lī seng-téng chòe chō͘-lí kàu-siū, m̄-koh bú seng-téng bú kah chin thiám-thâu, khai-sí ū pe̍h-thâu-mo͘ a.

Kán Li̍p-hong lāu-su koh ū kā góa kóng chi̍t ê chin tiōng-iàu ê koan-liām, i kā góa kóng, sî-kan kàu--a, tō ài chòe kai chòe ê tāi-chì, chit kù ōe góa ū kā kì tī sim-koa^{n} tóe. Só͘-í, tī seng-téng kiat-kó chhut-lâi chìn-chêng, góa khai-sí sin-chhéng kok-kho-hōe kè-e̍k. Lēng-gōa, Tē^{n} Liông-úi kàu-siū hit-chūn tī Kau-tāi, i hi-bōng góa ē-tàng khì Chheng-hôa kè-sio̍k tha̍k phok-sū-pan.

Góa khai-sí liân-lo̍k, pau-koah Tiu^{n} Chùn-sēng lāu-su, āu-lâi mā koh liân-lo̍k Lí Sek-kian lāu-su kah Tân Sìn-hi lāu-su. In piáu-sī hoan-gêng, m̄-koh kám-kah chòe Tâi-gú gián-kiù beh pi̍t-gia̍p ū khùn-lân. Chha-put-to tī chit ê sî-chūn, Ko Sêng-iām lāu-su pān chi̍t tiû^{n} Tâi-gú-bûn siong-koan ê gián-thó-hōe, hōe-āu chòe-tīn chia̍h-pn̄g ê sî, i kiò góa tn̂g-lâi Tâi-tāi Chu-sìn-hē tha̍k, tō chòe Tâi-gú-bûn hong-bīn. Góa tn̂g-khì siū^{n} chi̍t lé-pài, che kî-kan koh ùi Hoa-liân cháu-khì Koe-lâng chhiú khan chhiú, koh kah Ko lāu-su liân-lo̍k, khak-jīm, liáu-āu chôa^{n}-á koat--lo̍h-lâi, khai-sí lâi chún-pí.

2004 nî, chha-put-to kāng chi̍t ê sî-chūn, seng-téng thong-kè, phok-sú-pan lo̍k-chhú, Kok-kho-hōe kè-e̍k sin-chhéng--tio̍h, Tan-hông mā beh ji̍p Hù-sió. Ūi-tio̍h Gio̍k-lêng chhōa gín-á hong-piān, gún ùi Kiat-an poa^{n} lâi Bí-lūn-á, tī Hù-sió āu-piah-mn̂g bóe chi̍t keng a-phà-toh. Góa siū^{n}-kóng choan-sim tha̍k khah khòai pi̍t-gia̍p, só͘-í Tāi-hàn chit pêng sin-chheng liû-chit thêng-sin.

Ko lāu-su chit ê sî-chūn chhéng-tio̍h Tâi-bûn-kóan ê kè-e̍k, beh chòe Tâi-gú ê gú-im ha̍k-sêng, pún-lâi chhiá^{n} góa ê tông-o̍h Sēng-an tàu-tīn chòe, khó-sioh tòng bô chi̍t kò ge̍h, tō koh kā Sēng-an tiàu khì chòe Bioinfo, Ko lāu-su ê pún-tô͘. Pún-lâi koh chhiá^{n} Hank chòe-tīn lâi chòe Tâi-gú gú-im piān-sek, bô gî-gō͘ i āu-lâi bô thong-kè chu-keh-khó chit koan. Kè-e̍k ū chi̍t ê choan-jīm ê gia̍h, chhiá^{n} Kia̍t-ga̍k hia^{n} lâi tàu-sa^{n}-kāng. Góa ka-kī ê kè-e̍k leh ? Sui-jiân sin-chhéng--tio̍h, chóng--sī pó͘-chō͘ kim-gia̍h siu^{n} chió, só͘-í chú-chhî-jîn hùi poah chhut-lâi chhiá^{n} Lē-soat lâi Hoa-liân chòe kè-e̍k ê khang-khè.

Kám-siā lāu-pē lāu-bú, ūi-tio̍h góa beh tn̂g-khì Tâi-pak tha̍k-chu, in te̍k-pia̍t kā chhù koh hoan chi̍t piàn. Góa chi̍t lé-pài tòa Tâi-pak gō͘-kang, m̄-koh lāu-su kau-tài--lo̍h-lâi ê khang-khè li-li khok-khok, mā bô-hoat-tō͘ choan-sim tha̍k-chu, koh-khah hāi ê, Gio̍k-lêng ùi 11 ge̍h khai-sí, khó-lêng sī seng-o̍ah siu^{n} kín-tiu^{n}, liân-sòa nn̄g kò ge̍h ūi-tn̂g bô sòng-khòai. Só͘-í góa liû-chit thêng-sin ûi-chhî chi̍t ha̍k-kî, liáu-āu tō koh tn̂g-lâi kà-chu, thang kā lāu-su kóng góa bōe-tàng tiā^{n}-tiā^{n} tī Tâi-pak.

Chu-keh-khó sī chi̍t kiā^{n} khióng-pò͘ ê tāi-chì, pêng-kun chi̍t kài ū 1/3 ê tông-o̍h in-ūi bô thong-kè chu-keh-khó hông thòe-ha̍k, chhiū^{n} góa chit-khóan nî-hè khah tōa ê lāu ha̍k-seng, ap-le̍k iû-kî tōa. 2005 nî 3 ge̍h, tē it kho Complexity án-nóa kè ê ? In-ūi tè tio̍h tha̍k-chu-hōe ê kî-tiong kúi pái, kám-siā lī-hāi ê Ông Hông-lûn tông-o̍h. 2005 nî 10 ge̍h koh khó kè Algorithm kah AI, góa kau bōe chió ha̍k-hùi hioh-jo̍ah ta̍k lé-pài cháu Tâi-pak chham-ka tha̍k-chu-hōe. Ē-kì-e khó-chhì chêng ê lé-pâi, Lêng-ông hong-thai lâi, Kiat-an ê chhù, chhù-téng sià chúi lo̍h-lâi, Bí-lūn-á ê chhù, po-lê-thang-á phòa--khì, thêng-chúi nn̄g kang, iah bōe-hù chhú-lí, thang-á seng iōng chóa kô͘--khí-lâi, pau-ho̍k-á khóan--leh tō kín pià^{n} khì Tâi-pak chún-pī khó-chhì, sim-koa^{n}-thâu phi̍t-pho̍k-chhái. 2006 nî 3

gėh chòe-āu chi̍t kho Architecture bô kè, ap-lėk koh piàn tōa, nî kè 40, hiông-hiông kám-kah chiân-tô͘ bông-bông, 2006 nî 10 gėh chòe-āu chi̍t pái ê ki-hōe, chóng-sǹg ôan-sêng tē sì kho chu-keh-khó, ná-chhiūⁿ í-keng keng-lėk chi̍t-sì-lâng hiah-nī kú.

Thȧ̍k phok-sū-pan ê kî-kan, tú-tio̍h 3 pái gin-á in-ūi hì-iām tòa īⁿ, Gio̍k-lêng siōng-pan cháu bōe khui kha, só͘-í sī góa khì pīⁿ-īⁿ kò͘, ē-kì-ê ū chi̍t-piàn, chai-iáⁿ beh tòa īⁿ, kín cháu-khì tô͘-su-kóan chioh chi̍t pún “Chhèng-phàu, pīⁿ-khún kah kǹg-thih”, lī-iōng tòa tī pīⁿ-īⁿ ê sî-kan thȧ̍k ôan, che sī Khîn-hōaN hiaⁿ kài-siāu góa thȧ̍k ê. Chit tōaⁿ sî-kan, góa koh ū thȧ̍k “Tiūⁿ-niû sè-kài”, “chi̍t tháng ō-á”, “Hun-lân keng-iām”, “Tâi-ôan chhú-hun” … téng-téng ê chu, lâi pâi-kái sim-thâu ê ut-chut, tāi-pō͘-hūn sī tī hé-chhia téng thȧ̍k ôan ê. Koh ū chi̍t tōaⁿ sî-kan, chi̍t thàu-chá khí-lâi seng sńg chi̍t ê Sudoku, chiām-sî kā chòe bōe ôan ê khang-khè khǹg chi̍t piⁿ.

Pâi-kái ut-chut, góa koh ū chòe pȧ̍t-hāng, pau-koah khì chham-ka Google sió kang-khū pí-sài. Thiaⁿ Kán Li̍p-hong lāu-su kóng, góa ê Tâi-hôa sû-tián sió kang-khū ū ji̍p-ûi, chóng--sī, āu--lâi su hō͘ sǹg-miā ê sió kang-khū. Tâi-ôan-lâng ài sǹg-miā, khah iâⁿ kè o̍h bú-gú, si̍t-chāi chin bû-nāi. Kám-siā Lí Chì-kiông lāu-su, i chio gún khiâ khóng-bêng-chhia, 2008 nî 3 gėh góa ū tī 20 tiám-cheng í-lāi ôan-sêng 300 khí-lò, iân-lō͘ Seven bōe ê sio ê chia̍h-mi̍h lóng chhiúⁿ bô, hit àm thiaⁿ chhân-kap-á ê siaⁿ thiaⁿ kah pá, khiâ kah kha kiông-beh tn̄g--khì, kha-chhng o̍ah-beh thiàⁿ--sí, tńg--lâi koh sûi sán 3 kong-kun, m̄-koh tit-tio̍h chi̍t tiuⁿ chèng-su.[2] Kám-siā Tō͘ Chèng-sèng pō͘-tiúⁿ, i kā Kok-gú-bûn kēng-sài cheng-ka Tâi-gú pheng-im pí-sài, hō͘ góa ū ki-hōe tit-tio̍h siā-hōe-cho͘ tē it miâ,[3] i koh sak thui-tián bú-gú kia̍t-chhut kòng-hiàn-chióng, góa ū tio̍h-chióng, mā in-ūi tio̍h-chióng, hō͘ góa ê siau-sit ū ki-hōe khan tī Chū-iû sî-pò ê tōe-hng-pán.[4]

Kè-ėk ê pō͘-hūn, thȧ̍k phok-sū-pan ê chit tōaⁿ kî-kan, chin hó-ūn, chi̍p-hêng sì ê Kok-kho-hōe kè-ėk, kám-siā Kia̍t-gȧ̍k hiaⁿ, Tek-hôa kah Tōa-thâu-liân kā góa tàu chòe chin chōe khang-khè. Lēng-gōa, kám-siā Lī Heng-chhiong lāu-su, tī i ê

[2] http://ungian.pixnet.net/blog/post/16038270
[3] http://iug.csie.dahan.edu.tw/iug/Ungian/engu/2007/phengimpisai/phengimpisai.asp
[4] http://iug.csie. dahan.edu.tw/iug/Ungian/engu/2008/bugujit/bugujit.asp

chhui-chiàn hā, góa chi̍p-hêng chi̍t ê Tâi-bûn-kóan ê kè-e̍k, mā kám-siā Tiu^{n} Iām-hiàn lāu-su, i ê chhui-chiàn, hō͘ góa sin-chhéng tio̍h Kok-sú-kóan ê kè-e̍k, chòe Tâi-gú-bûn ūn-tōng hóng-tâm, kám-siā Bí-chhin, i thòe góa kā siōng khùn-lân ê hóng-tâm pō͘-hūn chhú-lí hó, Kok-sú-kóan chit pún chu, 2008 nî 4 ge̍h chhut-pan 800 pún, 8 ge̍h tō bōe ôan a. Chòe che bô pí siá phok-sū lūn-bûn khah khin-sang, ūi-tio̍h chit ê hóng-tâm, liân-sòa kúi-lō kò-ge̍h lóng òa^{n} khùn, kè-nî ê sî koh mo̍h kui thá̍h ê kó-kiā^{n} tih khòa^{n}.[5]

Mā kám-siā Há̍k-khiam hia^{n}, Khîn-hōa^{n} hia^{n}, Ûi-bûn, Gô͘ Chiau-sin i-su, Si-chong hia^{n} kah Sò͘-eng, in iau-chhiá^{n} góa chòe in ê ke-e̍k ê kiōng-tông chú-chhî-jîn, só͘-í góa ê keng-le̍k ē-tàng siá kah chiok súi-khùi ☺

Kám-siā Û Bêng-heng lāu-su, i kā góa iau-kó, hō͘ góa hoat-piáu góa ê tē it phi^{n} kok-chè kî-khan lūn-bûn, kám-siā Lâu Chiau-lîn kah Tân Pek-lîm lāu-su, in ê iau-kó sī tē jī phi^{n}. Kám-siā Tân Sìn-hi lāu-su, in-ūi i ê iau-chhiá^{n}, hō͘ góa ū ki-hōe chòe Chu-sìn jī há̍k-mn̂g ê kui-ōe úi-ôan, thê-chhut góa tùi pún-thó͘ gú-giân chu-sìn chhú-lí ê kiàn-gī.

Kám-siā Tāi-hàn ki-su̍t há̍k-ī^{n}, góa tī chiâ chia̍h thâu-lō͘, chit-chūn sī tē 10 tang, tī chiâ, góa khai-sí ē-tàng choan-sim lâi chòe Tâi-gú-bûn ê há̍k-su̍t gián-kiù, mā tī chiâ seng-téng chòe chō͘-lí kàu-siū, sui-jiân chit keng há̍k-hāu í-keng put-chí-á hûi-hiám a.

Kám-siā Tâi-tang tāi-há̍k Tâi-gú-só͘ ê há̍k-seng, sui-jiân hioh-jo̍ah ta̍k lé-pái cháu Tâi-tang put-chí-á sin-khó͘, m̄-koh in jīn-chin kah chun-tiōng lāu-su ê thāi-tō͘, hō͘ góa kám-siū tio̍h chòe lāu-su ê chun-giâm. Bo̍k-chiân ū 3 ê góa chí-tō ê há̍k-seng pit-gia̍p, hō͘ góa chin ū sêng-chiū-kám.

Kám-siā Chhit-chhi^{n}-thâm, Chhit-chhi^{n}-thâm m̄ sī thâm, sī tōa-hái, kè-khì chit tang, góa tiā^{n}-tiā^{n} khiâ khóng-bêng-chhia siong-hā-pan, khòa^{n} tōa-soa^{n} tōa-hái ê hó kéng-tì, koh kiam liān sin-thé.

Khó-lêng koh ū chi̍t kóa lâng góa bô kóng--tio̍h, m̄-sī lín bô tiōng-iàu, sī góa kì-tì bái.

Tē it pái ji̍p Tâi-tāi, sī 1984 nî khó tio̍h Chu-sìn-hē, tō͘-kè jîn-seng

[5] http://writtentaiwanesemovement.blogspot.com/2008/04/blog-post_16.html

cheng-chhái ê saⁿ tang gōa, kiàn-lı̍p kè-tа̍t-koan kah Tâi-ôan ì-sek. Tē jī pái jı̍p Tâi-tāi sī 1994 nî tī Tâi-tāi tiān-sòan tiong-sim chiа̍h thâu-lō͘, hit-chūn thé-hōe Tâi-ôan tōa-soaⁿ ê súi, mā ôan-sêng chiong-sin tāi-sū. Tē saⁿ pái jı̍p Tâi-tāi, tō-sī chit pái thа̍k phok-sū-pan, hi-bōng ē-tàng hō͘ góa chhē tio̍h khah ún-tēng ê thâu-lō͘.

Ka-têng, khang-khè, gián-kiù, hа̍k-giа̍p, … Nî-hè khah tōa, tāi-chì chōe, thа̍k-chu khí-lâi chin sńg sin-thé, chit sì tang pòaⁿ í-lâi, bа̍k-chiu ū ló-hoe, koh tio̍h pe̍h-lāi-chiong, chhùi-khí chhut būn-tôe, koaⁿ-kong-lêng chí-sò͘ chhèng kôan, siⁿ phê-chôa, kha-tóe mā ē thiàⁿ, kui sin-khu pīⁿ liáu-liáu. Góa kî-si̍t m̄ chai-iáⁿ thа̍k phok-sū-pan sī m̄-sī chèng-khak ê koat-tēng, chóng--sī, Tâi-gú-bûn chiâⁿ-chòe góa it-seng ê chì-giа̍p, che sī phah sí bô thòe ê.

Chòe-āu, beh kám-siā Tâi-gú-bûn-kài chōe-chōe ê pêng-iú, in-ūi ū tа̍k-ke ê phah-piàⁿ, góa chiah ū ki-hōe chòe chit ê tôe-bo̍k. Tâi-gú-bûn ūn-tōng ê lō͘ iah-koh chin kú-tn̂g, lán chòe-tīn kè-sio̍k kiâⁿ lo̍h-khì, ǹg-bāng Tâi-gú-bûn chhut-thâu-thiⁿ !

ps: Lūn-bûn kàu-sî ē chiūⁿ-bāng, chit phiⁿ siā-sû ē khǹg tī góa ê Blog.[6]

ps2: Acknowledgments ê chho͘-kó ū khǹg khì Tâi-gú-bāng, kám-siā Siau Pêng-tī lāu-su, Tiuⁿ Ho̍k-chû i-su, Ngô͘ Chiau-sin i-su, Tân Phek-siū bs, Lâm-hêng-hiaⁿ, Kim-êng lāu-su, Jîn-sek lāu-su, Khái--ko, Siok-hūi, Tsùn-tsiu ê liа̍h-pau kah chàn-siaⁿ. Chit pō͘-hūn siá liáu siōng sóng-khòai, bián chhin-chhiūⁿ siá lūn-bûn an-ne, tа̍k jī tа̍k kù ài chim-chiok.

ps3: 2009/1/16 sui-jiân thong-kè kháu-chhì, m̄-koh ài chun-chiàu kháu-chhì úi-ôan ê ì-kiàn siu-kái, in-ūi kin-nî kè-nî chá, hа̍k-hāu iau-kiû siōng òaⁿ 2/2 ài kau, siu-kái ê sî-kan bōe-hù, phah-sǹg iân chi̍t hа̍k-kî chiah thang pit-giа̍p, bô gî-gō͘ 1/20 siu-tio̍h hа̍k-hāu ê thong-ti, kóng kau lūn-bûn ê kî-hān kái-chòe 2/15. Che ke--chhut-lâi ê nn̄g lé-pài, ná-chhiūⁿ sī chiam-tùi góa ê tiâu-khóaⁿ. Hō͘ góa thang lī-iōng kè-nî hioh-khùn ê sî-kan kóaⁿ siu-kái ê khang-khè, kî-hān lāi ôan-sêng siu-kái, kiám la̍p chi̍t hа̍k-kî kùi-som-som ê chù-chheh-hùi. Kám-siā Tâi-tāi, mā kám-siā Gio̍k-lêng ê thé-thiap.

[6] http://ungian.pixnet.net/blog/post/25236012

摘要

台語是世界上重要的語言，可惜沒有受到應有的重視。在某些方面，台語文的特性與華文或英文相當不同。本論文主要討論台語文處理技術。

白話字（台語羅馬字）是台語文的重要書寫系統。我們先介紹白話字的字元編碼，提及白話字數字調號做爲不同白話字字元編碼的內部表示法。針對白話字文本搜尋，我們提出兩階段搜尋策略，並提出白話字音節近似搜尋的方法。我們還描述白話字顯示方法、白話字文字處理相關應用程式以及漢羅台語文斷詞方法。

我們提出以規則方法處理變調問題的演算法。先將每個台語詞翻成華語詞，找出其詞類標記訊息，以詞類標記和變調規則來決定變調後的聲調。我們實作出台語變調系統。此系統在訓練資料及測試資料分別達到 97.4%和 89.0%的變調正確率。

此外，我們提出詞類標記方法。我們先開發語詞對齊檢查程式將逐段對齊的兩種台語文本做語詞對齊，之後利用 HMM 機率模型挑選最適當的華語對應詞，再利用 MEMM 分類器挑選出其詞性標記。我們的方法達到 91.5%的正確率。

過去幾年，我們建立了一些有用的線上台語文工具。希望這些工具以及我們所做的初步研究成果，能讓台語文處理相關研究更加蓬勃發展。

關鍵詞： 台語文，變調，詞類標記，白話字，自然語言處理。

Tiah-iàu

Tâi-gú sī sè-kài siōng chin tiōng-iàu ê gú-giân, khó-sioh soah bô hông khòa^{n}-tiōng. Tī bo̍ chi̍t-kóa hong-bīn, Tâi-gú-bûn ê te̍k-sèng kah Hôa-bûn iah-sī Eng-bûn chin bô kāng-khóan. Pún lūn-bûn chú-iàu beh thó-lūn Tâi-gú-bûn chhú-lí ki-su̍t.

Pe̍h-ōe-jī (Tâi-gú Lô-má-jī, POJ) sī Tâi-gú-bûn ê tiōng-iàu su-siá hē-thóng. Gún seng kài-siāu POJ ê jī-gôan pian-bé, thê-chhut iōng POJ sò͘-jī tiāu-hō chiâ^{n}-chòe bô kâng POJ jī-gôan pian-bé ê lōe-pō͘ piáu-sī-hoat. Chiam-tùi PÔJ bun-pún chhiau-chhē, gún thê-chhut nn̄g kai-tōa^{n} chhiau-chhē chhek-lio̍k, koh thê-chhut POJ im-chiat lūi-sū chhiau-chhē ê hong-hoat. Gún koh ū kài-siāu POJ hián-sī ê hong-hoat, POJ ê bûn-jī chhú-lí siong-koan èng-iōng thêng-sek, kah Hàn-lô tn̄g-sû ê hong-hoat.

Gún thê-chhut iōng kui-chek hong-hoat chhú-lí piàn-tiāu būn-tôe ê ián-sòan-hoat. Seng kā múi chi̍t ê Tâi-gú sû hoan chòe Hôa-gú sû, chhē chhut i ê sû-lūi piau-kì, iōng sû-lūi piau-kì kah piàn-tiāu kui-chek lâi koat-tēng piàn-tiāu liáu-āu ê sia^{n}-tiāu. Gún khai-hoat chhut Tâi-gú piàn-tiāu hē-thóng. Chit ê hē-thóng tī hùn-liān chu-liāu kah chhì-giām chu-liāu hun-pia̍t ta̍t-kàu 97.4% kah 89.0% ê piàn-tiāu chèng-khak-lu̍t.

Lēng-gōa, gún thê-chhut sû-lūi piau-kì hong-hoat. Gún seng khai-hoat gú-sû tùi-chôe kiám-cha thêng-sek, kā chi̍t tōa^{n} chi̍t tōa^{n} tùi-chôe ê nn̄g-khóan Tâi-gú bûn-pún chòe gú-sû ê tùi-chôe, liáu-āu lī-iōng HMM ki-lu̍t bo͘-hêng kéng chhut siōng tàu-tah ê Hôa-gú tùi-èng-sû, koh lī-iōng MEMM hun-lūi-khì kéng chhut sû-lūi piau-kì. Gún ê hong-hoat ta̍t-kàu 91.5% ê chèng-khak-lu̍t.

Kè-khì kúi-lō tang, gún kiàn-li̍p chi̍t-kóa chin ū lo͘-ēng ê sòa^{n}-téng Tâi-gú-bûn kang-khū. Hi-bōng chiâ ê kang-khū kah gún só͘ chòe ê chho͘-pō͘ gián-kiù sêng-kó, thang hō͘ Tâi-gú-bûn chhú-lí siong-koan gián-kiù koh-khah heng.

Koan-kiān-sû : Tâi-gú-bûn, Piàn-tiāu, Sû-lūi piau-kì, Pe̍h-ōe-jī, Chū-jiân gú-giân chhú-lí.

Abstract

Taiwan Southern Min (Taiwanese) is an important language that has received very little attention in the world. The characteristics of written Taiwanese are quite different from Mandarin or English in some respects. This dissertation focuses on Taiwanese processing techniques.

POJ is an important Taiwanese script. We introduce the POJ character code and mention numbered POJ as the interchange code for various POJ encodings. Then, we propose a two-stage search strategy for a POJ text search, and propose a POJ syllable query expansion. We also describe the display method for POJ, POJ word processing utilities, and a word segmentation method for HR mixed script.

We propose a rule-based tone sandhi algorithm. We translated every word into Mandarin, and obtained the POS information. Using the POS data and tone sandhi rules, we then tagged each syllable with its post-sandhi tone marker. Finally, we implemented a Taiwanese tone sandhi processing system. Our system achieved accuracy rates of 97.4% and 89.0% with training and test data, respectively.

Additionally, we propose a POS tagging method. We developed a word alignment checker to assist with the word alignment of the two Taiwanese scripts, selected the most adequate Mandarin word using a Hidden Markov probabilistic model, and finally tagged the word using a Maximal Entropy Markov Model classifier. We achieved an accuracy rate of 91.5% in the Taiwanese POS tagging work.

Over the past several years, we have established some useful online written Taiwanese tools. Based on these tools and our preliminary research results, we hope that written Taiwanese processing related research can be promoted.

Keywords: Written Taiwanese, Tone Sandhi, POS Tagging, Peh-Oe-Ji, Natural Language Processing

Table of Contents

Preface i
Acknowledgments iv
摘要 xi
Abstract xiii
Abbreviations xxiii
Chapter 1 Introduction 1
1.1 Background 1
1.1.1 Language Population in Taiwan 1
1.1.2 Southern Min Language Population 2
1.1.3 Another Investigation: the Taiwan Southern Min Viewers 3
1.1.4 The Confusing Name of This Language 5
1.2 Different Types of Written Taiwanese Scripts 6
1.2.1 The Han Characters Script 7
1.2.2 The Romanized Scripts 9
1.2.3 The Han-Romanization Mixed Script 10
1.2.4 Other Scripts 10
1.2.5 Target Scripts in This Dissertation 11
1.3 Issues Related to Written Taiwanese Processing 11
1.4 Organization of This Dissertation 12
Chapter 2 Resources and Survey of Written Taiwanese Processing 19
2.1 Digital Resources for Written Taiwanese 19
2.1.1 Fonts 19
2.1.2 Dictionary 20
2.1.3 Text Corpora 23
2.1.4 Electronic Books 27
2.2 Survey of Written Taiwanese Processing Techniques 28
2.2.1 Input Method 28

2.2.2 Word Segmentatation 29
2.2.3 POS Tagging 30
2.2.4 Scripts Conversion 30
2.2.5 Text-to-Speech 30
2.2.6 Translation 33
2.2.7 Parsing 33
2.3 Summary 33
Chapter 3 Coding, I/O for POJ, and Text Processing 35
3.1 Character Code of POJ 35
3.2 Two Kinds of POJ Representation 39
3.3 Search Problem with POJ Text 41
3.3.1 Issues with POJ Text Search 41
3.3.2 Two-Stage Search Method: String Matching Then Filtering 42
3.3.3 Query Expansions: Toneless, Glottal Stop, Checked Syllable, and Vowel 44
3.3.4 Examples of Search Results 47
3.4 POJ Text Display 49
3.4.1 Issues with POJ Text Display 49
3.4.2 POJ and Numbered POJ Conversion Method 50
3.4.3 POJ Graph Display 52
3.4.4 Examples of Display Results 53
3.5 Some Text Processing Utilities for POJ 55
3.5.1 POJ Phoneme Segmentation and Spelling Checker 55
3.5.2 POJ Syllable/Word/Sentence Count 57
3.6 Word Segmentation for HR Mixed Script 58
Chapter 4 Tone Sandhi Problem and Algorithm 63
4.1 Tone Sandhi Problem of the Taiwanese Language 63
4.1.1 Types of the Taiwanese Language Tone Sandhi 64

4.1.2 Boundary of Tone Sandhi Group 68
4.2 Implementation of the Taiwanese Pronunciation System 68
4.2.1 System Diagram 68
4.2.2 Observation Data and Test Data 70
4.2.3 POS Tagging Set 71
4.2.4 Tone Sandhi Marks 73
4.3 Rule-based Tone Sandhi Algorithm 73
4.4 Results, Accuracy Rate and Discussion 78
4.4.1 Experiment Results 78
4.4.2 Accuracy Rate and Related Analysis 80
4.4.3 Discussion 83
4.5 Summary and Possible Direction 85
Chapter 5 POS Tagging Method 87
5.1 Problems of POS Tagging 87
5.2 POS Tagging Methods 88
5.2.1 Origin of the Corpus 89
5.2.2 Word for Word Alignment 89
5.2.3 Searching for the Corresponding Mandarin Candidate Words 90
5.2.4 Selecting the Best Mandarin Translation 91
5.2.5 Selecting the Most Appropriate POS According to the Corresponding Mandarin Word 92
5.3 Results 94
5.4 Error Analysis 99
5.4.1 Incorrect Corresponding Mandarin Word Selection 99
5.4.2 Absence of Appropriate Mandarin Words in the OTMD 100
5.4.3 Unknown Words from the Viewpoint of Mandarin 101
5.4.4 Propagation Error 101

5.4.5 Other Cases 101
5.4.6 Summary of Error Conditions 102
5.5 Discussion 103
5.5.1 Is Improvement Possible ? 103
5.5.2 Hyphen Problems, Distinction between Taiwanese and Mandarin 104
5.5.3 The Distinction between Different Eras or Different Genres 105
5.6 Summary 106
Chapter 6 Conclusion and Future Work 109
6.1 Our Contributions to Written Taiwanese Resources and Processing 109
6.2 Future Work and Prospects for Written Taiwanese Processing Research 112
Reference 117
Appendix 127
A.1 Brief Introduction to The Phoneme of Taiwanese 127
A.1.1 Initials 127
A.1.2 Vowels 128
A.1.3 Tones 129
A.1.4 Compared with Mandarin 130
A.2 Examples of Written Taiwanese 132
A.3 Terminologies 136
A.4 Webpages Made by Author 138
A.5 Differences between POJ and TL 139

List of Tables

Table 1 - 1 Living Languages in Taiwan 1
Table 1 - 2 Southern Min Language Populations around the World 2
Table 1 - 3 Visits to the Taiwan Southern Min Contents Website 4
Table 1 - 4 The Names of “Taiwan Southern Min” 6
Table 3 - 1 POJ Unicode Encoding 36
Table 3 - 2 POJ and Numbered POJ Synopsis 40
Table 3 - 3 Example of Taiwanese Article in 3 Scripts 40
Table 3 - 4 Toneless Search for “hoe-chhia” 48
Table 3 - 5 Checked Syllable Search for “cha” 48
Table 3 - 6 Vowel Search for “uiN” 49
Table 4 - 1 The Taiwanese Tone Sandhi Phenomena 67
Table 4 - 2 Observation Data Sources 70
Table 4 - 3 Test Data Sources 71
Table 4 - 4 POS Classes 72
Table 4 - 5 Tone Sandhi Marks 73
Table 4 - 6 Tone Sandhi Marking Algorithm 73
Table 4 - 7 Number of Errors and Accuracy Rate for Each Paragraph 80
Table 4 - 8 Affected and Accurately Affected Syllables of Each Rule 81
Table 4 - 9 Number of Dominant Rule, Accurate Dominate Rule and Accuracy Rate 82
Table 4 - 10 Additional Rules to Obtain Higher Accuracy Rate 83
Table 4 - 11 Error Conditions and Possible Solutions 84
Table 5 - 1 Test Data List 95
Table 5 - 2 Tagging Accuracy Rate for the Test Data 96
Table 5 - 3 Example of POS Tagging Result 97
Table 5 - 4 The Incorrect Mandarin Words Selected and Their Respective POS 100

Table 5 - 5 Errors Caused by the Absence of Appropriate Mandarin Words in the OTMD..... 100
Table 5 - 6 Unknown Words from the Viewpoint of Mandarin 101
Table 5 - 7 The Reasons for the POS Tagging Errors..... 102
Table 5 - 8 Tagging Accuracy Rates for Different Genres..... 106
Table 5 - 9 Tagging Accuracy Rates for Different Eras..... 106
Table A - 1 Consonants..... 127
Table A - 2 Single Vowels..... 128
Table A - 3 Compound Vowels..... 128
Table A - 4 Nasal and Glottal..... 128
Table A - 5 Nasal Finals and Final Stops..... 129
Table A - 6 Syllabic Consonants..... 129
Table A - 7 Tones..... 130
Table A - 8 Terminologies..... 136
Table A - 9 Differences between POJ and TL..... 139

List of Figures

Fig 3 - 1 POJ Text Match Algorithm (Target: Word) 43
Fig 3 - 2 POJ Toneless Search Algorithm (Target: Syllable) 45
Fig 3 - 3 POJ Checked Syllable Search Algorithm (Target: Syllable) 46
Fig 3 - 4 POJ Vowel Search Algorithm (Data: Syllable) 47
Fig 3 - 5 POJ to Numbered POJ Algorithm 50
Fig 3 - 6 Numbered POJ to POJ Algorithm 51
Fig 3 - 7 Numbered POJ to POJ Graph Algorithm 53
Fig 3 - 8 Unicode Display of POJ 54
Fig 3 - 9 Graph Display of POJ 54
Fig 3 - 10 Check If a Legal POJ Syllable Algorithm 56
Fig 3 - 11 Result from Online Syllable/Word/Sentence Count System for POJ 58
Fig 3 - 12 Backward Maximal Matching Algorithm for HR Mixed Script 59
Fig 4 - 1 Taiwanese Tone Sandhi System Diagram 69
Fig 5 - 1 Taiwanese Language POS Tagging System Architecture Diagram 89
Fig A - 1 Han script (Taiwanese Folk Song) 132
Fig A - 2 Han script (Taiwanese Textbook in Japanese-Ruled Period) 133
Fig A - 3 POJ script (Taiwan Prefectural City Church News) 134
Fig A - 4 HR Mixed Script (Taiwanese Writing Forum) 135

Abbreviations

Abbrev.	Taiwanese (POJ)	English	Mandarin
CCA	Bûn-kiàn-hōe	Council for Cultural Affairs	文建會
CCTPLD	Tâi-ôan POJ bûn-ha̍k chu-liāu so͘-chi̍p	The Collection and Cataloging of Taiwanese POJ Literature Data	台灣白話字文學資料蒐集
CED	Tiong-bûn tiān-chú sû-tián	Chinese electronic dictionary	中文電子辭典
CKIP	Tiong-bûn sû tì-sek-kho͘ sió-cho͘	Chinese Knowledge and Information Processing	中文詞知識庫小組
CWSTS	Tiong-bûn tn̄g-sû kap Sû- lūi Phiau-kì hē-thóng	Chinese Word Segmentation and Tagging System	中文斷詞及詞類標記系統
DADWT	Tâi-gú-bûn so͘-ūi tián-chông chu-liāu-kho͘	Digital Archive Database for Written Taiwanese	台語文數位典藏資料庫
HMM	Ún Markov mô͘-hêng	Hidden Makov Model	隱馬可夫模型
LDC	Gú-giân Chu-liāu Tiong-sim	Language Data Consortium	語言資料中心
MEMM	Siōng tōa lōan-tō͘ Markov mô͘-hêng	Maximal Entropy Markov Model	最大熵馬可夫模型
MOE	Kàu-io̍k-pō͘	Ministry of Education	教育部
NMTL	Tâi-ôan bûn-ha̍k-kóan	National Museum of Taiwan Literature	台灣文學館
NSC	Kok-kho-hōe	National Science Council	國科會
OTCS	Tâi-gú-bûn gú-sû kiám-sek hē-thóng	Online Taiwanese Concordancer System	台語文語詞檢索系統
OTJDTT	Tâi-Ji̍t tōa sû-tián Tâi-gú e̍k-pún	Online Taiwanese- Japanese Dictionary with Taiwanese Translation	台日大辭典台語譯本

Abbrev.	Taiwanese (POJ)	English	Mandarin
OTMD	Tâi-hôa sòa^{n}-téng sû-tián	Online Taiwanese- Mandarin Dictionary	台華線上辭典
OTSD	Tâi-gú sòa^{n}-téng jī-tián	Online Taiwanese Syllable Dictionary	台語線上字典
POJ	Pe̍h-ōe-jī	Taiwanese vernacular writing	白話字
POS	Sû-sèng	Part of speech	詞性
SMHLA	Bân-Kheh-gú tián-chông	Southern Min and Hakka Language Archive”	閩客語典藏
TAICORP	Tâi-ôan jî-tông gú-liāu-khò͘	Taiwan Child Language Corpus	台灣兒童語料庫
THMLW	Tâi-gú kap Kheh-gú hiān-tāi bûn-ha̍k choan-tôe bāng-chām	Taiwanese and Hakka Modern Literature Website	台語及客語現代文學專題網站
TL	Tâi-ôan Bân-lâm-gú Lô-má-jī pheng-im hē-thóng hong-àn	Taiwan Southern Min Lō-má-jī phonetic scheme	臺灣閩南語羅馬字拼音方案
TVLA	Tâi-ôan Pe̍h-ōe-jī bûn-hiàn chu-liāu-kóan	Taiwanese Vernacular Literature Archive	台灣白話字文獻資料館