

# Thai Word Segmentation Web Service

The screenshot shows the THAISEMANTIES website with a blue header bar. The header contains the logo 'THAISEMANTIES' and the subtext 'Free Thai language resources and services'. Below the header is a navigation bar with links: หน้าหลัก (Home), User Profile, Swath Word Segment, Orchid POS Tagger, and Web Service. The main content area has a light blue background. On the left, there's a 'Navigation' sidebar with links to Home, Swath Word Segment, Orchid POS Tagger, and Web Service. Below that is an 'Account' section with a welcome message 'Welcome, seksan.' and a 'Profile' link. The main content area starts with a breadcrumb 'หน้าหลัก > Service'. It then lists 'JSON HTTP Web Service' and provides an example input for the SWATH method: 'SWATH {api\_key: 'YOUR API KEY', 'method': 'SWATH', 'params': ['unicode strings'], }'. There's also a 'Support method' section.

Seksan Poltree ([seksan.poltree@gmail.com](mailto:seksan.poltree@gmail.com))  
Asst. Prof. Kanda Saikaew ([krunapon@kku.ac.th](mailto:krunapon@kku.ac.th))  
Department of Computer Engineering  
Faculty of Engineering  
Khon Kaen University

# Agenda

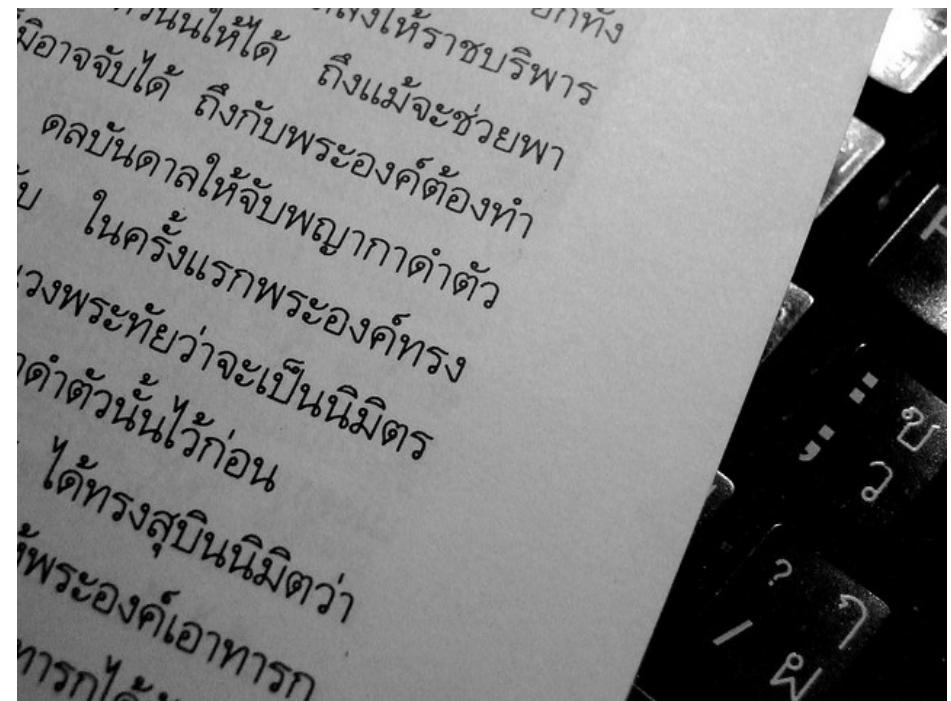
- Thai vs English text processing
- Current Thai Software and Service
- Why segmentation web service
- System Overview
- Web Application Example
- Provided Service Methods
- Comparing Service vs TLEX
- Conclusion and Future work

# Current Thai Software and Service

Resource	description	Licensing
libthai	Segmentation software + word list corpus Maximal Matching	GNU LGPL
SWATH	Segmentation software + word list corpus Maximal matching/ longest matching	GNU GPL
ORCHID	Thai Part-Of-Speech tagged corpus	NECTEC (BSD-like)
BEST	Thai segmentation solution corpus	NECTEC (BSD-like)
TLeX Service	SOAP Web service Conditional Random Field technique	Free to use

# Thai vs English in Text Processing

- Extract Thai Words?
  - no boundaries
  - no delimiters
- Word Segmentation is a classical issue
- Need word and sentences segmentation

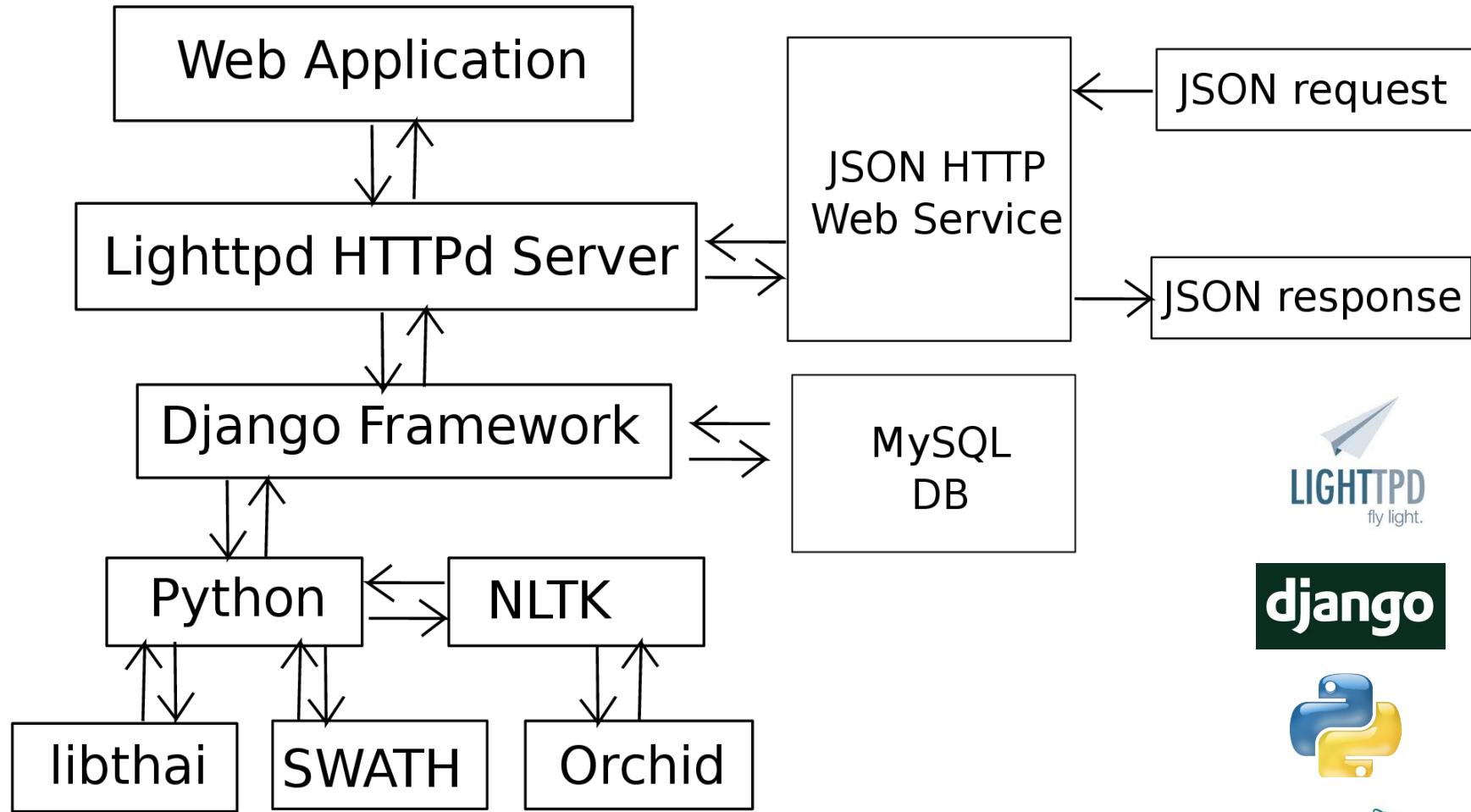


# Why Segmentation Web Service

- Increasing of web application and services
- Reducing user learning time of segmentation algorithms
- Make use of existing Thai language resources



# System Overview



# Web Application : SWATH

## Swath Word Segmentation Service

Input Sentence	Segmented Output
<p>เพื่อเปิดกระดูกกระซิบและข้อของปลาเรา จะเห็นเหวี่อกอยู่ ภายใน บนต้านหน้าของกระดูกโครงจะเหวี่อกจะมีส่วนที่ยื่น ออกมากเป็นชี้เรียวยาวหรือเป็นศูนย์ ส่วนนี้เราเรียกว่า ชี้ เหวี่อก ปลาที่กินพืชน้ำ เช่น กินสาหร่าย ชี้เหวี่อกของ ปลาเหล่านี้จะสั้นและมีจำนวนน้อย ส่วนปลาจ้าพวกที่กิน แพลงก์ตอน เป็นอาหาร ชี้เหวี่อกยาวเรียวและมีเป็น จำนวนมากสาหรับปลาบางชนิด เช่น ปลาญู ชี้เหวี่อก แหลมชี้รังแทกแข็งและออกไประดิษ์ หัวนี้เพื่อช่วยเพิ่ม สมรรถภาพให้แก่ปลาในการกรองอาหารขนาดเล็กจาก น้ำ ปลากินเนื้อริบส่วนใหญ่เป็นอาหาร ชี้เหวี่อกอาจจะ</p>	<p>เพื่อ   เปิด   กระดูก   กระซิบ   และ   ข้อ   ของ   ปลา   เรา   จะ   เห็น   เหวี่อก   อยู่   ภายใน   บน   ต้านหน้า   ของ   กระดูก โครง   เหวี่อก   จะ   มี   ส่วน   ที่   ยื่น   ออก   มาก   เป็น   ชี้   เรียว   ยาว   หรือ   เป็น   ศูนย์   ล ส่วน   นี้   เรา   เรียกว่า   ชี้   เหวี่อก   ปลา   ที่   กิน   พืช   น้ำ   เช่น   กิน   สาหร่าย   ชี้   เหวี่อก   ของ   ปลา   เหล่านี้   จะ   สั้น   และ มี   จำนวน   น้อย   ส่วน   ปลา   จ้าพวก   ที่   กิน   แพลงก์ ตอน   เป็น   อาหาร   ชี้   เหวี่อก   ยาว   เรียว   และ   มี   เป็น   จำนวน   มาก   สาหรับ   ปลา   บาง   ชนิด   เช่น   ปลาญู   ชี้   เหวี่อก   แหลม   ชี้   รัง   แทก   แข็ง   และ   ออก   ไประดิษ์  </p>
<input type="button" value="Segment"/>	<input type="button" value="ถ้า"/>

<http://www.thaisemantics.org/service/swath/index>

# Web Application : ORCHID

## Orchid Part of Speech Service

Input Sentence	Segmented and Tagged Output
เมื่อเปิดกระดูกกรทุ่งแมกมของปลาเราจะเห็นเหวือกอยู่ ภายใน บนต้านหน้าของกระดูก็โครง เหวือกจะมีส่วนที่ยื่น ออกมาก เป็นชี้เรียวขาวหรือเป็นตุ่ม ส่วนนี้เราเรียกว่า ชี้ เหวือก ปลาที่กินพืชน้ำ เช่น กินสาหร่าย ชี้เหวือกของ ปลาเหล่านี้จะสีน้ำเงินและมีจำนวนน้อย ส่วนปลาจ้าวක์ที่กิน แพลงก์ตอน เป็นอาหาร ชี้เหวือกขาวเรียวและมีเป็น จำนวนมากสาหรับปลาบางชนิด เช่น ปลาๆ ชี้เหวือก แต่จะชี้ยังแคกและงอกไปอีก ทั้งนี้เพื่อช่วยเพิ่ม สมรรถภาพให้แก่ปลาในการกรองอาหารขนาดเล็กจาก น้ำ ปลากินเนื้อหรือล่าเหยื่อเป็นอาหาร ชี้เหวือกอาจจะ เป็น	เมื่อ/JSBR   เปิด/VACT   กระดูก/NCMN   กระทุ่ง แมกม/UNK   ออก/RPRE   ปลา/NCMN   เรา/PPRS   ฉะ/XVBM   เห็น/VSTA   เหวือก/UNK   อยู่/XVAE   ภายใน/RPRE   <space>/PUNC   บน/RPRE   ต้าน หน้า/NCMN   ของ/RPRE   กระดูก/NCMN   โครง/UNK   เหวือก/UNK   ฉะฝี/UNK   ส่วน/NCMN   ที่/PREL   ยืน/UNK   ออก/XVAE   ผา/XVAE   เป็น/VSTA   ชี้/NCMN   เรียว/UNK   ขาว/VATT   หรือ/JCRG   เป็น/VSTA   ตุ่ม/UNK   <space>/PUNC   ส่วน/NCMN   นี้/DDAC   เรา/PPRS   เรียกว่า/VACT   <space>/PUNC

<http://www.thaisemantics.org/service/orchid/index>

# Current Provided Service Methods

	Request Format	Response Format
<b>SWATH</b>	<pre>api_key': 'YOUR API KEY', 'method': 'ORCHID', 'params': [['list','PoS'],['OF','PoS'],['list','PoS'], ['list','PoS']], }</pre>	<pre>{"status": 0, "result": ['list','of', 'segmented', 'words'], }</pre>
<b>ORCHID</b>	<pre>{'api_key': 'YOUR API KEY', 'method': 'ORCHID', 'params': [['list','PoS'],['OF','PoS'],['list','PoS'], ['list','PoS']], }</pre>	<pre>{"status": 0, "result": [list of tagged', 'words'], }</pre>
<b>Wrong KEY</b>	<pre>{ 'api_key': '', 'method': 'ORCHID', 'params': ['unicode strings'], }</pre>	<pre>{"status": 1, "result": ["Wrong API key."]}</pre>
<b>Wrong JSON</b>	{unknown or malform json format}	<pre>{"status": -1, "result": [ "Unkown request"]}</pre>

# Register to get Free API Key

- Using Facebook account instead of legacy registration



**Request for Permission**

ThaiSemantics ขออนุญาตทำสิ่งต่อไปนี้:

เข้าถึงข้อมูลส่วนตัวที่มีฐานของฉัน  
เพิ่มชื่อ, รูปภาพ, โปรไฟล์, เพศ, เครื่องข่าย, ยูสเซอร์ ไอดี, รายชื่อเพื่อน, และข้อมูลอื่นๆ ที่ได้แสดงต่อสาธารณะ

ส่งอีเมลให้ฉัน  
ThaiSemantics สามารถส่งอีเมลล์ถึงฉันได้โดยตรงที่ sekasan.poltree@gmail.com · เป็นยืน

Post to Facebook as me  
ThaiSemantics อาจใช้โพสต์ข้อความสถานะ, บันทึก, ภาพถ่ายและวิดีโอในนามของฉัน

รายงานและพิจารณา

ลงชื่อเข้าใช้ในชื่อของ Seksan Poltree · ออกจากรหัสบัน

**อนุญาต** **ไม่อนุญาต**

- Re-generated your API Key on demand

Account

Welcome, sekasan.

- Profile
- Logout

Profile details

API-KEY: 4b6736c86a0af25a9cb0c5497848ec64779f8bdea7252d1a886a498f48085C

Re-Hash

# Why REST, not SOAP Service?

- **REST** : REpresentational State Transfer
  - Simple, Lightweight
  - But Lack of Standard
- **SOAP** : Simple Object Access Protocol
  - XML based, Schema, Standard
  - Need more bandwidth, Higher round trip time Latency
  - No complex schema description need for segmentation
  - REST is more suitable!



# Why JSON not XML

- **XML** : eXtensible Markup Language
  - Self Descriptive language
  - Mark up overhead
- **JSON** : JavaScript Object Notation
  - Use simple brackets and notations
  - Suitable for simple transfer data
- No complex schema description need for segmentation , JSON is more suitable!



# Comparing Service with TLeX

## Original BEST data

เนื่องจากทั้งประเด็นเรื่อง|"ความไม่เป็นธรรม"|และเรื่อง|"คู่ต่างเป็นประเด็นที่ใหญ่และซับซ้อน|จึงส่งผลให้การสะกดต่างๆ เช่น |เส้นอต่างๆ|ที่เสnonในบทนำ|ดังกล่าวมีลักษณะนี้| ข้อเสนอต่างๆ|ที่เสnonในบทนำ|ดังกล่าวนี้| จึงเป็นสิ่งที่ต้องการให้ท่านผู้อ่านได้นำไปใช้คิด|ถกเถียงกันต่อไป

## SWATH Output

เนื่องจากทั้งประเด็นเรื่อง|"ความไม่เป็นธรรม"|และเรื่อง|"คู่ต่างเป็นประเด็นที่ใหญ่และซับซ้อน|จึงส่งผลให้การสะกดต่างๆ เช่น |เส้นอต่างๆ|ที่เสnonในบทนำ|ดังกล่าวมีลักษณะนี้| ข้อเสนอต่างๆ|ที่เสnonในบทนำ|ดังกล่าวนี้| จึงเป็นสิ่งที่ต้องการให้ท่านผู้อ่านได้นำไปใช้คิด|ถกเถียงกันต่อไป

## TLex Output

เนื่องจากทั้งประเด็นเรื่อง|"ความไม่เป็นธรรม"|และเรื่อง|"คู่ต่างเป็นประเด็นที่ใหญ่และซับซ้อน|จึงส่งผลให้การสะกดต่างๆ เช่น |เส้นอต่างๆ|ที่เสnonในบทนำ|ดังกล่าวมีลักษณะนี้| ข้อเสนอต่างๆ|ที่เสnonในบทนำ|ดังกล่าวนี้| จึงเป็นสิ่งที่ต้องการให้ท่านผู้อ่านได้นำไปใช้คิด|ถกเถียงกันต่อไป

- Using BEST corpora as test data
- Create simple script and call each service
- TLEX and SWATH use difference method and implementation
- Just prove of concept

# Evaluation Result

AVERAGE SERVICE CALLING USAGE TIME

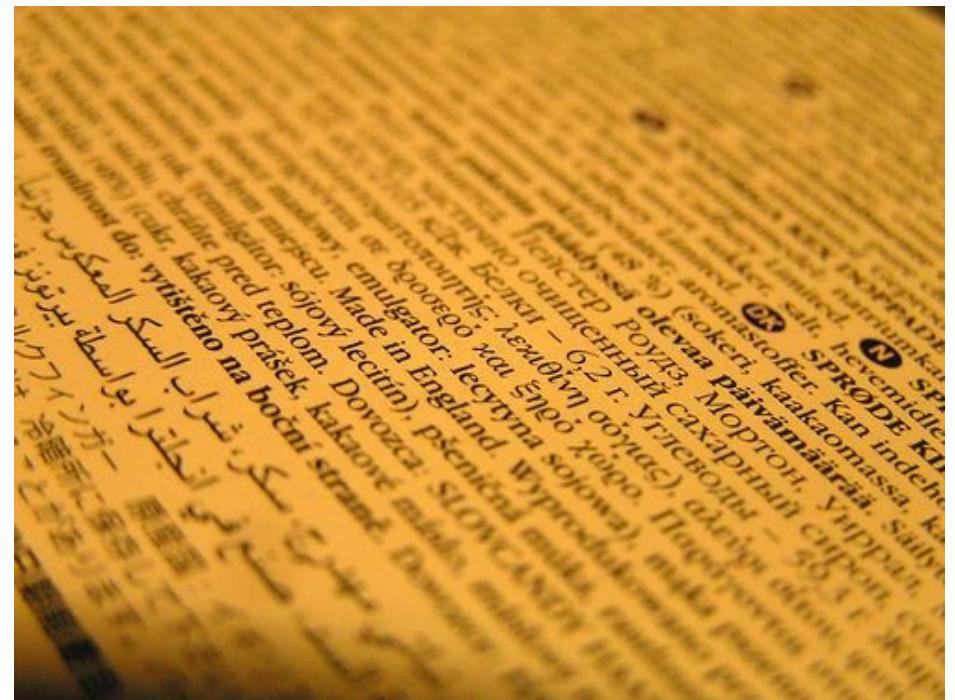
Calling Service	real	user	sys
SWATH	1m39.293s	0m5.172s	0m1.176s
ORCHID	4m50.377s	0m11.077s	0m1.212s
TLexs	3m17.045s	0m6.632s	0m1.048s

EVALUATION RESULTS

Calling Service	Precision	Recall	F-Score
SWATH	85.93	77.62	81.57
TLexs	95.65	97.45	96.54

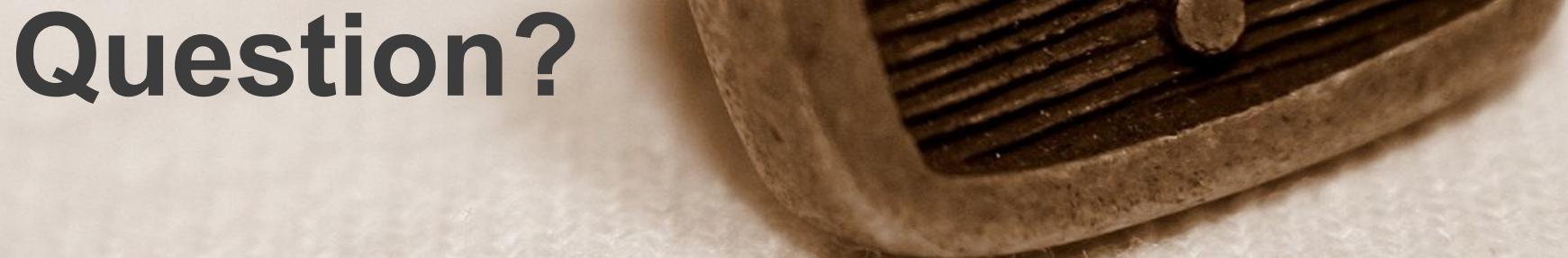
# Conclusion and Future work

- Create Segmentation and POS-Tagger application and services
- Create Free JSON REST Web Service
- <http://www.thaisemantics.org>
- Comparing with existing TLeX SOAP web service to prove of concept
- Include more method and corpus in the future
- Using facebook account instead of registration



# References

- [1] T. Karoonboonyanan, C. Silpa-Anan, P. Kiatisevi, P.Veerathanabutr and V. Ampornaramveth, "libthai Library". Available at: <http://linux.thai.net/projects/libthai>.
- [2] P. Charoenpornsawat, "SWATH (Smart Word Analysis for THai)". Available at: <http://www.cs.cmu.edu/~paisarn/software.html>.
- [3] T. Karoonboonyanan, "swath 0.4.1 Released". Available at: <http://linux.thai.net/svn/software/swath>.
- [4] The Royal Institute of Thailand 2525, "Thai dictionary words from the royal institute of Thailand 2525". Available at: <http://thailang.nectec.or.th/>.
- [5] V. Sornlertlamvanich, T.Charoenporn and H.Isahara, "ORCHID: Thai Part-Of-Speech Tagged Corpus". National Electronics and Computer Technology Center. Technical Report: TR-NECTEC-1997-001, 1997.
- [6] National Electronics and Computer Technology Center (NECTEC), "BEST Corpus". Available at: <http://thailang.nectec.or.th/best/>.
- [7] C. Haruechaiyasak and S. Kongyoung, "TLex: Thai Lexeme Analyser Based on the Conditional Random Fields", *Proc. 8th International Symposium on Natural Language Processing*, 2009.
- [8] National Electronics and Computer Technology Center (NECTEC), "TLex". Available at: <http://sansarn.com/tlex/>.
- [9] National Electronics and Computer Technology Center (NECTEC), "TLexs". Available at: <http://www.sansarn.com/WSeg/wsdl/BnSeg.wsdl>.
- [10] K. Toutanova and C. D. Manning, "Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger". *Proc. the Joint SIG-DAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 63-70, 2000.
- [11] D. Roth and D. Zelenko, "Part of Speech Tagging Using a Network of Linear Separators", *The 17th International Conference on Computational Linguistics* (1998), pp. 1136-1142, 1998.
- [12] The World Wide Web Consortium. "Extensible Markup Language (XML) 1.0 (Fifth Edition)". Available at : <http://www.w3.org/TR/REC-xml>.
- [13] The World Wide Web Consortium. "SOAP Version 1.2 Part 1: Messaging Framework (Second Edition)". Available at : <http://www.w3.org/TR/soap12-part1/>.
- [14] R. T. Fielding. "Representational State Transfer (REST)". Available at : [http://www.ics.uci.edu/~fielding/pubs/dissertation/rest\\_arch\\_style.htm](http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm).
- [15] Ecma International, "Introducing JSON". Available at: <http://www.json.org/>
- [16] Ecma International, "Standard ECMA-262 5.1 Edition / June 2011 : ECMAScript Language Specification". Available at: <http://www.ecma-international.org/publications/files/ecma-st/ECMA-262.pdf>
- [17] Wikipedia, "Comparison of web browsers". Available at: [http://en.wikipedia.org/wiki/Comparison\\_of\\_web\\_browsers](http://en.wikipedia.org/wiki/Comparison_of_web_browsers)
- [18] C. Haruechaiyasak, S. Kongyoung and M. N. Dailey. "A Comparative Study on Thai Word Segmentation Approaches", *Proc. 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology* (2008), pp.125-128, 2008.
- [19] G. van Rossum, "Python tutorial, Technical Report CS-R9526", *Centrum voor Wiskunde en Informatica (CWI)*, Amsterdam, May, 1995.
- [20] S. Bird and E. Loper, "NLTK: The Natural Language Toolkit", *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pp. 62-69, 2002.
- [21] Django Software Foundation, "The Django framework". Available at : <https://www.djangoproject.com/>.
- [22] "LightHTTPD fly light". Available at : <http://www.lighttpd.net/>.
- [23] C. D. Manning, P. Raghavan and H. Schutze. "An Introduction to Information Retrieval", *Cambridge University Press, Cambridge, England*, pp.155, 2009.
- [24] G. Mulligan and D. Gračanin. "A Comparison of SOAP And REST Implementation of a Service Based Interaction Independence Middleware Framework", *Proceedings of the Winter Simulation Conference*, pp.1423-1432, 2009.
- [25] G. Wang. "Improving Data Transmission in Web Applications via the Translation between XML and JSON", *Third International Conference on Communications and Mobile Computing*, pp.182-185, 2011.



# Question?