A Large-scale Lexical Semantic Knowledge-base of Chinese [*]

Wang Hui
Department of Chinese Studies
National University of Singapore
Singapore 117570
chswh@nus.edu.sg

Yu Shiwen
Institute of Computational Linguistics
Peking University
Beijing 100871, China
yusw@pku.edu.cn

Abstract

The Semantic Knowledge-base of Contemporary Chinese (SKCC) is a large scale Chinese semantic resource developed by the Institute of Computational Linguistics of Peking University. It provides a large amount of semantic information such as semantic hierarchy and collocation features for 66,539 Chinese words and their English counterparts. Its POS and semantic classification represent the latest progress in Chinese linguistics and language engineering. The descriptions of semantic attributes are fairly thorough, comprehensive and authoritative. The main work in this paper is to introduce the outline of SKCC, and establish a multi-level WSD model based on it. The results indicate that the SCK is effective for word sense disambiguation in Chinese and are likely to be important for general NLP.

Key words: semantic knowledge-base, word sense disambiguation (WSD), lexical semantics, Chinese language processing.

1 Introduction

Semantic resources play an important role in many areas of Natural Language Processing (NLP). The Institute of Computational Linguistics (ICL) of Peking University has been engaged in research and development of *the Semantic Knowledge-base of Contemporary Chinese* (SKCC) in the last eight years. This lexicon-building project was collaboration with the Institute of Computing Technology, Chinese Academy of Sciences during 1994-1998, and resulted in a machine-readable bilingual lexicon suitable for use with Machine Translation applications, which contained a fairly complete characterization of the semantic classification, valence specifications and collocation properties for 49 thousands Chinese words and their English counterparts [1].

Since 2001, the further development of SKCC has been co-conducted by ICL and Chinese Department of Peking University. At present, SKCC has made great progress. Not only is the scale extended to 66,539 entries, but also the quality has been immensely improved. The semantic classification in the updated edition of SKCC is the embodiment of the very latest progress in Chinese linguistics and language engineering, while the semantic descriptions are comprehensive and thorough. It can provide rich lexical semantic information for various NLP applications.

[*] Supported by China National Fundamental Research Program (973) (PN: G1998030507-4 & G1998030507-1).

2 Outline of SKCC

2.1 Scale and Structure

SKCC consists of one general database and six sub-databases. The general database contains shared attributes of all the 66,539 entries, while the sub-databases provide detailed descriptions of the distinctive semantic attributes associated with the parts of speech (POS). For example, the verb database has 16 attribute fields, noun database and adjective database has 15 attribute fields respectively (see table 1).

Database Name	Entries	Attribute fields	Attribute value
nouns	38,478	15	576,555
verbs	21,142	16	338,272
adjective	5,577	15	83,655
pronouns	236	15	3,540
adverbs	997	11	10,967
numerals	109	11	1,199
General	66,539	8	532,312
Total	133,078	**91**	**1,546,500**

Table 1 Scale of SKCC

All of the six sub-databases can be linked to the general database through four key fields, namely ENTRY, POS, HOMOMORPHISM and SENSE. As a result, the son knots can inherit all information from their father knots (Figure 1).

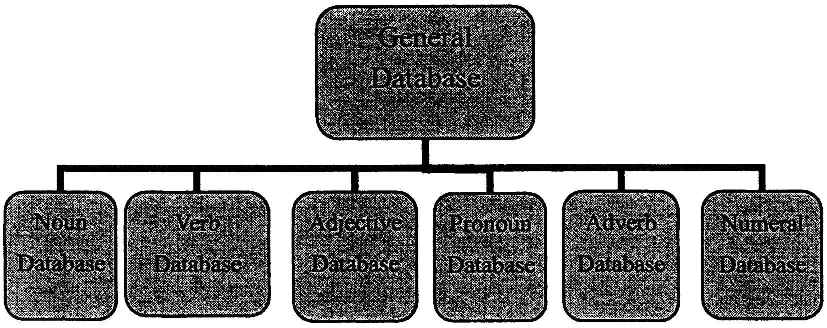
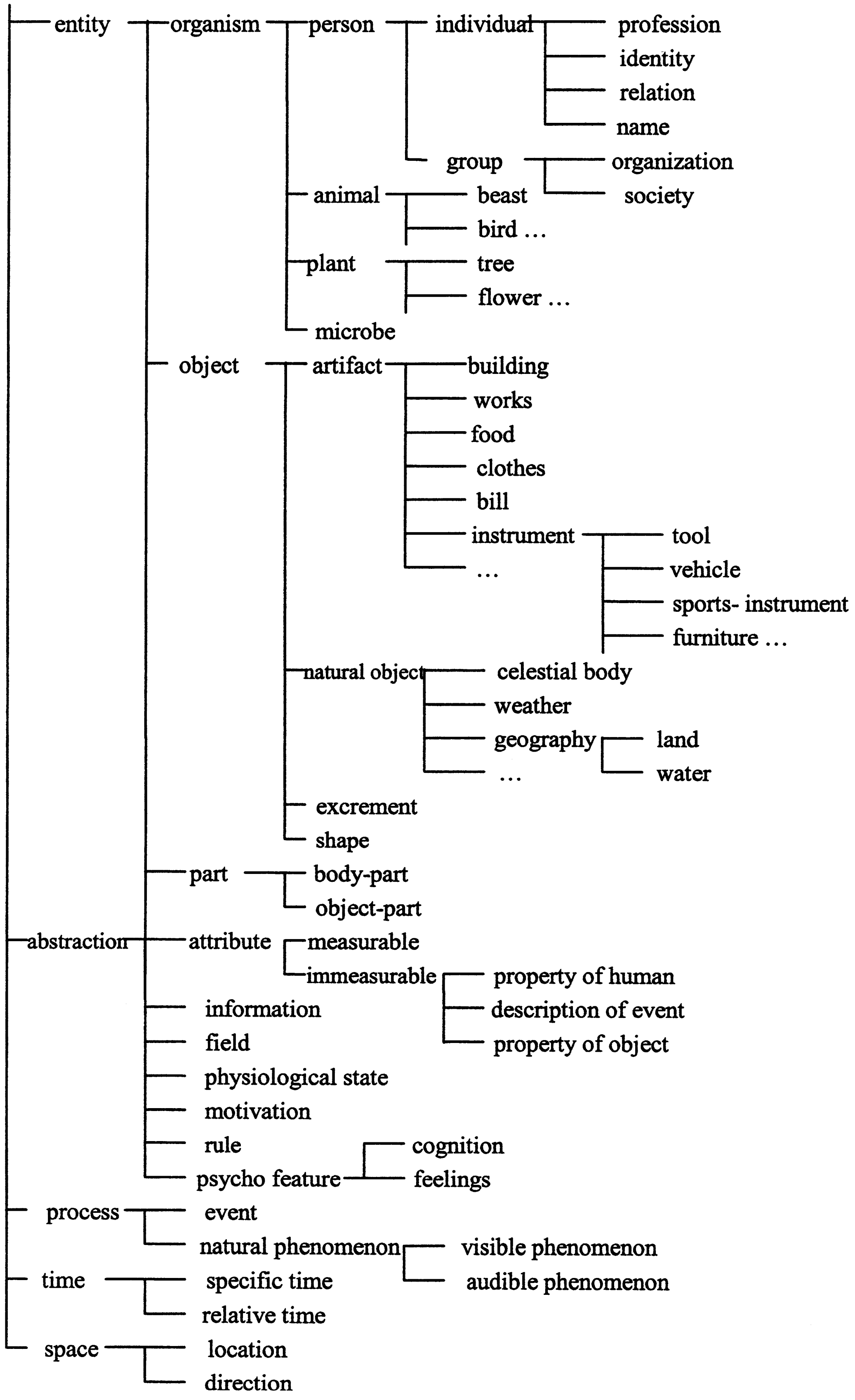


Figure 1 Main structure of SKCC

2.2 Semantic Hierarchy

One of the most outstanding characteristics of SKCC is that its semantic hierarchy is based on grammatical analysis, rather than merely on general knowledge (as illustrated in Figure 2). This classification system represents the latest progress in Chinese semantics. It is very useful for NLP applications[2], as well as compatible with various semantic resources, such as Wordnet[3], Chinese concept dictionary (CCD)[4], HowNet[5] etc. Currently, the classification of all of the 66,539 entries has already been completed.

(1) Nouns



- entity
  - organism
    - person
      - individual
        - profession
        - identity
        - relation
        - name
      - group
        - organization
        - society
    - animal
      - beast
      - bird ...
    - plant
      - tree
      - flower ...
    - microbe
  - object
    - artifact
      - building
      - works
      - food
      - clothes
      - bill
      - instrument
        - tool
        - vehicle
        - sports- instrument
        - furniture ...
      - ...
    - natural object
      - celestial body
      - weather
      - geography
        - land
        - water
      - ...
    - excrement
    - shape
  - part
    - body-part
    - object-part
- abstraction
  - attribute
    - measurable
    - immeasurable
      - property of human
      - description of event
      - property of object
  - information
  - field
  - physiological state
  - motivation
  - rule
  - psycho feature
    - cognition
    - feelings
- process
  - event
  - natural phenomenon
    - visible phenomenon
    - audible phenomenon
- time
  - specific time
  - relative time
- space
  - location
  - direction

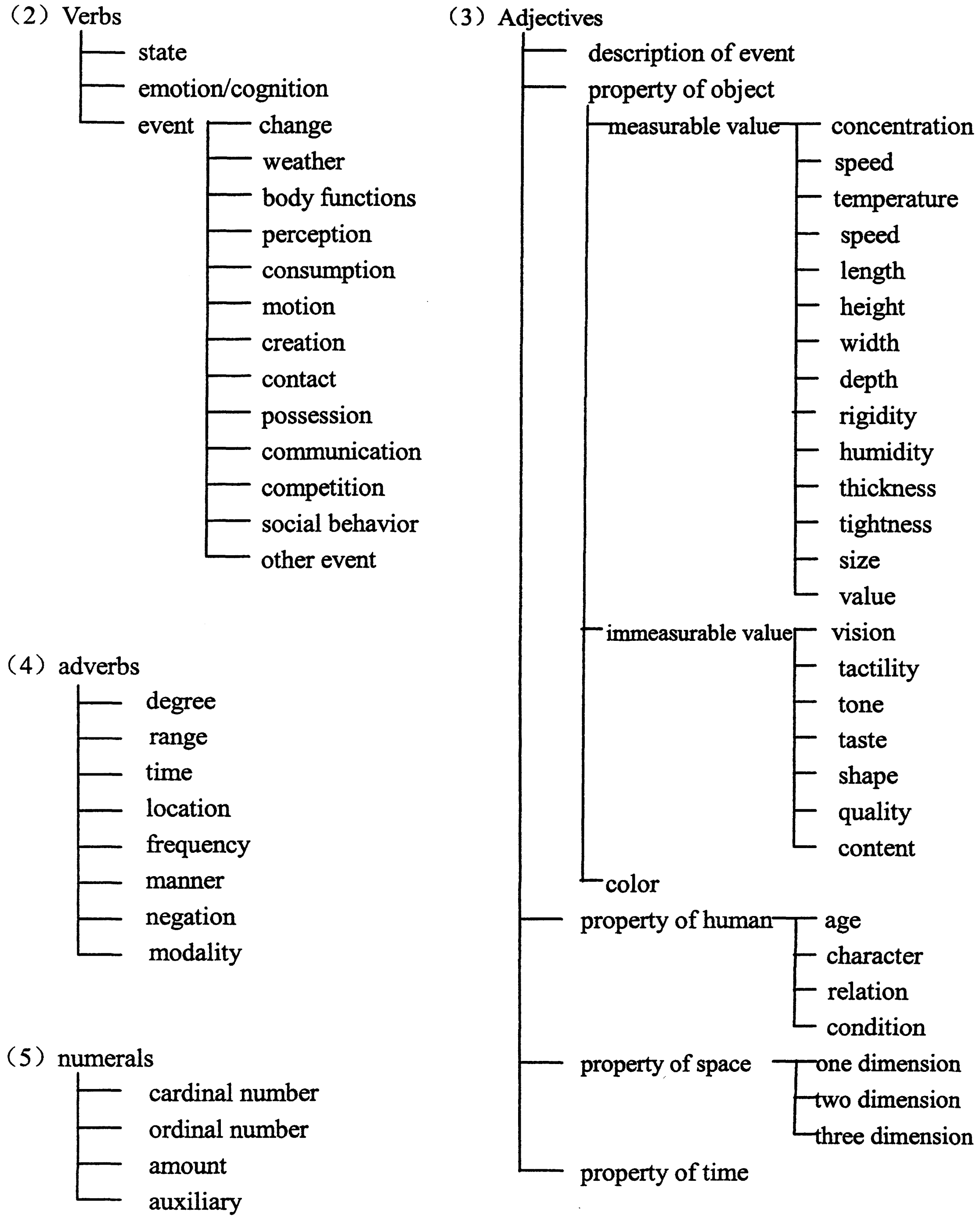


Figure 2 Semantic hierarchies in SKCC

2.3 Comprehensive Semantic Descriptions

There is close correlation between lexical meaning and its distribution. Oriented to MT and Natural Language Understanding applications, SKCC can provide detailed semantic description and collocation behavior that in many cases is likely to be uniquely associated with a single sense. For example, following attribute fields have already been filled with values in the verb database (see table 2).

FIELD	VALUE
ENTRY	Commonly used Chinese word or idiom phrase
PRONUNCIATION	Chinese Pinyin with tones such as "chi3zi5" for "尺子"(ruler)
PART OF SPEECH	POS tagging of per word or idiom
SUB-CATEGORY	Sub-category tagging of per word or idiom
POSs	All POS tagging of per word
HOMOMORPHISM	Homograph number
SENSE	Sense number of per polysemous word
DEFINITION	Sense definition
SEMANTIC CATEGOR	Semantic categories of per word or idiom.　A word can be tagged with two or more semantic categories. For instance, the noun "青菜" (greengrocery) belong to "plant | food" categories.
VALENCE	Valence number of each entry.　For example,"咳嗽"(cough) is a one-valence verb，while "吃"(eat) is a two-valence one, "给"(give) is three-valence.
AGENT	Actor of action or motion.
OBJECT	Object of action.
DATIVE	Beneficiary or suffer of action.
TRANSLATION	English counterpart of per word or idiom.
ECAT	POS tagging of per English word or phrase.
ILLUSTRATIONS	Corpus-derived example sentences showing authentic contexts of a word or idiom.

Table 2 Semantic attributes in the verb database of SKCC

To sum up, the above attributes fall into five kinds of information below:

(1) Basic information of entry, such as vocabulary item, part of speech, sub-category, homograph and pronunciation;

(2) Descriptions of word meaning, including sense number, definition, and semantic categories;

(3) Semantic valence, thematic roles and combinatorial properties for per words; this is the most important part of SKCC and especially useful for WSD and lexical semantics research;

(4) English translation and its POS tagging. If a Chinese word has two or more English counterparts, it will be regarded as different entries respectively, and the collocation information will also be given in relevant fields. This can significantly improve the quality of Chinese-English MT system.

(5) Corpus-derived authentic examples of a word in context, showing how it is used, how phrases are formed around it, and so on.

3 Application in WSD

As a large-scale lexical knowledge base, SKCC combines the features of many of the other resources commonly exploited in NLP work: it includes definitions and English translations for individual senses of words within it, as in a bilingual dictionary; it organizes lexical concepts into a conceptual hierarchy, like a thesaurus; and it includes other links among words according to several semantic relations, including semantic role, collocation information etc. As such it currently provides the broadest set of lexical information in a single resource. The kind of information recorded and made available through SKCC is of a type usable for various NLP applications, including machine translation, automatic abstraction, information retrieval, hypertext navigation, thematic analysis, and text processing.

In this section, we shall focus on the automatic disambiguation of Chinese word senses involving SKCC since it is most troublesome, and essential for all the above NLP applications [6].

3.1 Determination of the polysemous words and homographs

In general terms, word sense disambiguation (WSD) task necessarily involves two steps: (1) the determination of all the polysemous words and homographs in the text or discourse; and (2) a means to assign each occurrence of a word to the appropriate sense.

Step (1) can be easily accomplished by reliance on SKCC. Firstly, each entry denotes one single sense of per word in SKCC. Thus, if a word has two or more senses, it will be regard as different entries, and the "SENSE" field will be filled with different number (as "菜"in table 3).

ENTRY	POS	SENSE	DEFINITION	TRANSLATION	ILLUSTRATIONS
菜	n	1	vegetable	vegetable	种～(grow vegetables) | 野～(potherb)
菜	n	2	cooked vegetable, egg, fish, meat...etc.	dish	荤～(meat or fish) | 四～一汤(four dishes and a bowl of soup)

Table 3 Two senses of Chinese noun "菜"

Secondly, SKCC marked all of the homographs in "HOMOMORPHISM" field, such as two verbs "看"with different pronunciation in table 4.

ENTRY	PRONUNCIATION	HOMOMORPHISM	DEFINITION	Translation
看	Kan4	A	see; watch; look at	see
看	Kan1	B	look after; take care of	look after

Table 4 Homographs in SKCC

Therefore, if either of the "SENSE" and "HOMOMORPHISM" fields is filled with value in SKCC, the entry must be a polysemous word or homograph.

3.2 WSD based on semantic categories

The senses of most Chinese polysemous words and homographs belong to different semantic categories, and have different syntagmatic features in context [7]. SKCC gives detailed description of such information in “AGENT” and/or ”OBJECT” fields as illustrated in table 5 below.

ENTRY	POS	DEFINITION	SENSE	SEMANTIC CLASS	AGENT	TRANSLATI N
清淡	a	(of food, drink, smell) light; weak	1	taste	food | drink| plant	light
清淡	a	(of business) slack	2	condition	“business”	slack

Table 5 Polysemous adjectives in SKCC

Based on the above description, the target word “清淡” in following POS-tagged text can be accurately disambiguated:

[1] 一/m 杯/q 清淡/a 的/u 龙井茶/n (A cup of light Longjing tea)

[2] 农忙时/t 进城/v 的/u 人/n 不/d 多/a ， 生意/n 比较/d 清淡/a。

(When the season is busy, few farmers go to town and the business is rather slack)

In sentence[1], the word modified by “清淡” is the noun“茶”(tea) , which is a kind of “drink”; while the word “清淡” in sentence [2] is a predicate of “business”. According to the different values in ”AGENT” field, it is easy to judge that these two “清淡” belong to two semantic categories, viz. the former is “light”, and the latter is ”slack”.

3.3 WSD based on collocation information

As for the polysemous words or homographs belonging to the same semantic category, the difference between them usually manifests at the collocation level. According to a study in cognitive science, people often disambiguate word sense using only a few other words in a given context (frequently only one additional word) [8]. Thus, the relationships between one word and others can be effectively used to resolve ambiguity. For example, Chinese verb “找” has two senses: one is “寻找” (look for) and the other is “退还” (give change). Only when the verb co-occurs with the noun “钱” (money), it can be interpreted as “give change”; Otherwise, it means “look for” (see table 6).

ENTRY	HOMOMORPHISM	DEFINITION	AGENT	OBJECT	DATIVE	Translation
找	A	look for; seek try to find;	person	entity		look for
找	B	give change	person	“money”	person	give change

Table 6 Different senses of verb “找”

According to table 6, the verb ”找” in sentence [1] below must be “look for”, because its object is “人” (person), a kind of “entity”; while “找”in sentence [2] has two objects, namely, indirect object “我” (me) and direct object ”钱”(money). Thus, its meaning is “give change”.

[1]他们/r 将/d 出去/v 找/v 人/n。(They will go out to look for sb.)
[2]售货员/n 还/d 没有/d 找/v 我/r 钱/n 呢。（The seller has not given change to me）

By making full use of SKCC and a large scale POS-tagged corpus of Chinese, a multi-levels WSD model is developed and has already been used in a Chinese-English MT application.

4 Conclusion

SKCC is a well-structured Chinese-English bilingual semantic resource, as described in the paper, it has more than 66,000 Chinese words and their English counterparts classified, and the accurate description of about 1.5 million attributes further enriched the abundance of lexical semantic knowledge. It not only provides a deductive system of word meaning and valuable semantic knowledge for Chinese language processing, but also has great theoretical significance in lexical semantics and computational lexicography research.

Acknowledgement

We appreciate all the members participated in SKCC project, especially Prof. Lu Jianming, Mr. Li Kangnian and Dr. Chang Baobao. The blithesome collaboration with Dr. Ying Chenjin and Mr. Guo qingjian from Chinese Department is memorable for all of us. Lastly, thanks our colleagues and friends for their kindly discussion with the authors.

References

[1] Wang Hui, Zhan Weidong, Liu Qun. 1998. "Design of Semantic Dictionary of Modern Chinese". *Proceedings from 1998 International Conference on Chinese Information Processing*. Beijing: Tsinghua University Press. pp361-367.

[2] Zhan Weidong, Liu Qun. 1997. "The important role of semantic classification in Chinese-English MT". Language Engineering. Tsinghua University Press. 286-291.

[3] Christiane Fellbaum. 1998. WordNet: an electronic lexical database. Mass: MIT Press.

[4] Yu Jiangsheng, Yu Shiwen. 2002. "Structure and Design of CCD". Chinese Information Processing. 16 (4): 12-20.

[5] Dong Zhendong, Dong Qiang. "Hownet". http:// www.keenage.com.

[6] Ide, Nancy; Jean Véronis. 1998. "Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art", *Computational Linguistics*. Vol.24, No.1. pp1-40

[7] Wang Hui. 2002. "Chinese Word Sense Disambiguation in Machine Translation". *Proceedings from Chinese National Symposium on Machine Translation.* Beijing: Publishing House of Electronics Industry. pp.34-43.

[8] Choueka, Y. and S. Lusignan, 1983. "A Connectionist Scheme for Modeling Word Sense Disambiguation," *Cognition and Brain Theory*. 6 (1). pp.89-120.