Measure words are measurably different from sortal classifiers

Yamei Wang and **Géraldine Walther**
George Mason University
{ywang78, gwalthe}@gmu.edu

Abstract

Nominal classifiers categorize nouns based on salient semantic properties. Past studies have long debated whether sortal classifiers (related to intrinsic semantic noun features) and mensural classifiers (related to quantity) should be considered as the same grammatical category. Suggested diagnostic tests rely on functional and distributional criteria, typically evaluated in terms of isolated example sentences obtained through elicitation. This paper offers a systematic re-evaluation of this long-standing question: using 981,076 nominal phrases from a 489 MB dependency-parsed word corpus, corresponding extracted contextual word embeddings from a Chinese BERT model, and information-theoretic measures of mutual information, we show that mensural classifiers can be distributionally and functionally distinguished from sortal classifiers justifying the existence of distinct syntactic categories for mensural and sortal classifiers. Our study also entails broader implications for the typological study of classifier systems.

1 Introduction

Classifier systems constitute a major feature of East and South-East Asian languages (Li, 2013). Classifiers categorize referent nouns based on salient semantic features such as humanness, animacy, shape, or others (Aikhenvald and Mihas, 2019). In Mandarin, classifiers are obligatory when a noun is preceded by a number, a demonstrative, or a quantifier (Li and Thompson, 1989). For example, the classifier (in bold font) cannot be omitted in the following examples from Li (2013): 两 *liǎng* *(**个** *gè*) 学生 *xuéshēng* 'two students', 这 *zhè* *(**种** *zhǒng*) 动物 *dòngwù* 'this kind of animal', 每 *měi* *(**本** *běn*) 书 *shū* 'every book'. In other contexts, however, classifiers can be optional. In addition to *sortal classifiers*, which categorize nouns in terms of intrinsic semantic features, classifier systems also include *mensural classifiers* (or *measure words*), that are related to noun quantity. Table 1 lists a few common classifiers in Mandarin Chinese.

While sortal classifiers, like **张** *zhāng*, are typically associated with nouns displaying specific intrinsic semantic features, e.g., flat properties such as for the noun 地图 *dìtú* 'map', mensural classifiers like **组** *zǔ* 'group', **斤** *jīn* 'half kilogram', or **美元** *měiyuán* 'US Dollar' are usually characterized as being less restricted by the semantics of the nouns they combine with. In this paper, mensural classifiers like **组** *zǔ* 'group' will be referred as quantity, those like **斤** *jīn* 'half kilogram' and **美元** *měiyuán* 'US Dollar' will be referred as measurement and currency units respectively. In addition, Dryer et al. (2005) and Li (2013) indicate that sortal classifiers tend to be combined with countable nouns (e.g., 三 *sān* **本** *běn* 书 *shū* 'three books', 三 *sān* **只** *zhī* 碗 *wǎn* 'three bowls') while mensural classifiers refer to quantities of mass nouns (or "nouns with low countability") such as 三 *sān* **箱** *xiāng* 水 *shuǐ* 'three boxes of water' and 三 *sān* **斤** *jīn* 米 *mǐ* 'three half-kilograms of rice'. However those distinctions are not systematic: countable nouns can also be modified by mensural classifiers and mass nouns by sortal classifiers. For instance, the countable noun, 书 *shū* 'book' can be found in a nominal phrase such as 三 *sān* **箱** *xiāng* 书 *shū* 'three boxes of books', and the mass noun, 米 *mǐ* 'rice' can be modified by a sortal classifier in 三 *sān* **粒** *lì* 米 *mǐ* 'three grains of rice'.

Given their apparent similarities and differences, typological and general linguistic studies have long debated whether sortal and mensural classifiers should be considered as the same syntactic category (e.g., Lyons, 1977; Li and Thompson, 1989) or two different categories (e.g., Nguyen, 2004; Singhapreecha, 2001; Her and Hsieh, 2010). In these studies, suggested diagnostic tests rely on functional and distributional criteria, typically evaluated in terms of isolated example sentences

Proceedings of the Seventh International Conference on Dependency Linguistics (Depling, GURT/SyntaxFest 2023), pages 81 - 90
March 9-12, 2023 ©2023 Association for Computational Linguistics

Determiner	Classifier	Noun	Translation
一 *yī* 'one'	张 *zhāng* 'sortal classifier'	地图 *dìtú* 'map'	one map
这 *zhè* 'this'	组 *zǔ* 'group'	照片 *zhàopiàn* 'photo'	this group of photos
十二 *shíèr* 'twelve'	斤 *jīn* 'half kilogram'	米 *mǐ* 'rice'	six kilograms of rice
一亿 *yíyì* '100 million'	美元 *měiyuán* 'US Dollar'	公司 *gōngsī* 'company'	a 100 million US Dollar company

Table 1: Nominal phrases extracted from the Leipzig corpus of Mandarin Chinese (Goldhahn et al., 2012) using the CoreNLP Parser(Chen and Manning, 2014). The examples show noun phrases including the **sortal classifier** 张 *zhāng* and three measure words for **quantities** 组 *zǔ* 'group', **measurements** 斤 *jīn* 'half kilogram', and **currencies** *měiyuán* 'US Dollar'.

obtained through elicitation. We address this question in a more systematic and empirical way using data from large Mandarin corpora. We compare the distribution of sortal and mensural classifiers in terms of their contextual word representations and their function in terms of contribution to noun predictability. The idea that classifiers can be used to enhance the predictability of upcoming nouns is based on a study by Dye et al. (2017, 2018). The authors show that gendered determiners in German, which also partition the language's nouns into classes (masculine, feminine, and neuter), serve the communicative function of efficiently reducing the entropy of upcoming nouns in context. Similarly, our study adopts a communicative perspective on noun class partitioning and evaluates sortal vs. mensural classifiers in terms of their respective contribution to noun entropy reduction. If sortal and mensural classifiers prove to be distributed differently or to show differences in their degree of reducing the entropy of upcoming nouns, we will be able to successfully conclude that they can be considered separate syntactic categories. Otherwise, they would be better analyzed as two types within the same category.

Our study is based on 981,076 manually validated noun phrases extracted from a 489MB corpus of Mandarin Chinese that is part of the Leipzig Corpora Collection (Goldhahn et al., 2012), an open access collection of pre-cleaned data. We preparsed the data using the CoreNLP Chinese dependency parser (Chen and Manning, 2014). Our results show that mensural classifiers can be distributionally and functionally distinguished from sortal classifiers. Additionally two traditional subtypes of mensural classifiers (i.e., measurement and currency units) emerge as distinct from the other mensural classifier subtype (which we will refer to as *quantity*).

2 Measuring categorial differences

The goal of our study is to quantitatively evaluate whether distributional and functional properties of words traditionally labeled *sortal* vs. *mensural classifiers* suggest that they constitute a single or two separate syntactic categories in Mandarin Chinese. Based on 981,076 manually validated noun phrases extracted from a 489MB corpus, we analyze the syntactic distributions of sortal and mensural classifiers, as well as the differences in their communicative function for natural language use. We used contextual word embeddings as a measure of classifier distributions and mutual information (MI) (Cover and Thomas, 2012) as a measure of their contribution to facilitating noun predictability.

2.1 Data

We downloaded three of the 1M sentence corpora of Mandarin Chinese from the Leipzig Corpora Collection (Goldhahn et al., 2012)[1] and normalized the data by converting all Chinese characters into simplified Chinese using the Open Chinese Convert software.[2] We then applied the CoreNLP Chinese dependency parser (Chen and Manning, 2014) to our dataset. We used the dependency information to extract all complete nominal phrases containing nouns, classifiers, and other dependents such as determiners and adjectives, as well as the frequencies of all nouns and classifiers. 91 out of 1,079,190 nominal phrases were removed from the

[1]The types of corpora are 2007-2009 news, 2011 newscrawl, and 2015 China web: https://wortschatz.uni-leipzig.de/en/download/Chinese.

[2]https://github.com/BYVoid/OpenCC

data due to their unusual length (more than 35 characters)[3]. A sample of the remaining extracted nominal phrases is shown in Table 2.

Manual validation of the data revealed that despite the Mandarin Chinese CoreNLP parser's overall good performance[4] reported in Chen and Manning (2014), a significant proportion of words had been erroneously labeled as classifiers. We manually validated all 1,577 word types identified as classifiers by the parser. After excluding tokens containing symbols ('县、区、乡、'), non Chinese ('ま') or invalid characters ('\ue997'), numbers ('二九'), dialectal expressions ('拨儿'), and other similar cases, we were able to retain 315 classifier types. Excluded cases are listed in table 3. We labeled them as either sortal or mensural classifiers following the classification suggested in Chao (1965)'s reference grammar. Not all 315 classifier types are listed in Chao (1965). For those not listed in the grammar, we inferred the labels applying compatible classification criteria. As a result, 55 classifiers were labeled as sortal classifiers and 260 as mensural classifiers. We further categorized mensural classifiers into one of three subcategories: quantities (148), measurements (86), and currency (86) (see table 1 for examples of sortals, quantities, measurements, and currencies). The complete list of the classifiers with their corresponding labels is indicated in table 4.

Almost all discarded classifiers were hapaxes, such that at the end of the validation process, we were still left with 981,076 noun phrases out of the original 1,079,190. We analyzed the remaining classifiers in terms of their distributional properties (represented by contextual word embeddings) and functional properties (measured as mutual information (MI) (Cover and Thomas, 2012)).

Distributional information was obtained using the pre-trained Chinese BERT model from Hugging Face.[5] We extracted contextual word embeddings for all retained classifiers. Embeddings were based on the last-hidden state, where most of the contextual information is encoded.

Contextual word embeddings were adopted in order to be able to distinguish between tokens used as classifiers vs. identical tokens representing other parts of speech (e.g., 桶 *tǒng* 'bucket' represents a quantity in the phrase 一 *yī* 桶 *tǒng* 水 *shuǐ* 'one bucket of water' but is a noun the phrase 黄色 *huángsè* 的 *de* 桶 *tǒng* 'the yellow bucket'). Since the model returns one embedding per Chinese character, we were forced to discard classifiers represented by multi-character units[6]. This however only marginally changed our overall proportions for the two categories that lie at the core of this study: sortal classifiers vs. generic measure words. Overall, this step affected our four categories in the following way: sortal classifier tokens: 0% removed; generic measure word tokens: 1.4% removed; measurement tokens: 18.8% removed; and currency tokens: 99.3% removed. Because of their very specific use and homogeneous meanings, measurements and currency units are not usually considered contentious in the debate as to whether sortal and mensural classifiers constitute a unique or two separate categories. The removal of multi-character units from the currency units and measurement categories only removed a very small proportion of our overall classifier set and did not interfere with our main concern regarding the classification of sortal classifiers and generic measure words. At the end of this data cleaning process, our dataset contained 500,987 word embeddings corresponding to 221 distinct classifiers.

All noun frequencies (type: 324,920 and token: 27,596,565), frequencies of classifiers (type: 315 and token: 981,076), and their corresponding nouns (noun type: 45,159 and token: 981,076) in the retained nominal phrases were used to calculate the overall entropy of nouns and the Mutual Information between classifiers and their head nouns.

2.2 Method

Syntactic categories are commonly defined by the distribution and the function of the elements they contain. Words belonging to the same category are expected to display identical or similar distributional properties and functions. Our goal in this paper was to apply quantifiable measures of distribution and function to classifier tokens in order to objectively compare the distributional and functional properties of the types commonly identified

[3] Extracted phrases of more than 35 characters were judged to be abnormal by the author who is a native speaker. Given its small proportion, the removal of 91 our of 1,079,190 does not have a significant impact on the overall distributions.

[4] The unlabeled attachment score (UAS) and labeled attachment score (LAS) reported for the test dataset by Chen and Manning (2014) are 83.9% and 82.4% respectively. UAS indicates the percentage of words that have been assigned the correct head, and LAS shows the percentage of words that have been assigned the correct head and label.

[5] https://huggingface.co/bert-base-chinese

[6] Classifiers corresponding to multiple vectors.

Determiner	Classifier	Noun	Phrase
三 *sān* 'three'	部 *bù* sortal	片约 *piānyuē* 'film appointment'	三部片约 'three shooting sessions'
这 *zhè* 'this'	支 *zhī* sortal	团体 *tuántǐ* 'group'	这支出道 12 年的团体 'this 12-year-old group'
本 *běn* 'this'	周 *zhōu* 'week'	新闻 *xīnwén* 'news'	本周台湾旅游新闻 'Taiwan travel news of this week'

Table 2: Sampled nominal phrases extracted from the Leipzig corpus of Mandarin Chinese (Goldhahn et al., 2012) using CoreNLP Chinese Parser (Chen and Manning, 2014). The nominal phrase 三部片约 sān bù piānyuē 'three film shooting sessions' only contains a determiner, a classifier, and a noun. In addition to the determiner, classifier, and noun, the other two phrases also contain other modifying elements.

Discarded types	Examples
symbols	'II', '·', '〖'
characters with symbols	'县、区、乡、', '日圆、入不敷出'
invalid characters	'\ue997', '\ue08d
foreign characters	'ま', '４Ｇ'
numbers	'二九', '陆仟捌佰零'
combinations of numbers and symbols	'二·一六', '八〇八二六〇'
combinations of classifier + noun	'号楼', '吨钢'
combinations of noun + classifier	'人次','人份'
combinations of classifier + classifier	'吨级', '架次'
reduplicated classifiers	'盘盘', '首首'
verbal classifiers	'下','遍','次'
dialectal phrases	'拨儿','斗子'
phrases with typos	'豪米','届'
words not convertible into simplified Chinese	'場'
meaningless phrases	'圈共','岔起', '服轨'
words that cannot be classifiers	'跺', '啸', '恒星','烧烤', '富二代', '菩萨摩诃'

Table 3: Listed criteria used to manually validate parsed classifiers by using the CoreNLP parser (Chen and Manning, 2014)

as either sortal or mensural classifiers in the literature.

2.2.1 Exploring distributions

In order to evaluate the (dis)similarity between the distributions of our four classifier types, we compared their contextual word embeddings extracted from the pre-trained Chinese BERT model for all 221 single-character classifier types of our dataset. The embeddings produced by the model correspond to vectors with 768 dimensions for each token. We used the Uniform Manifold Approximation and Projection algorithm (UMAP) developed by McInnes et al. (2018) to perform high dimensionality reduction in order to better evaluate the (dis)similarity between the distributions of our four classifier types. UMAP maintains separability of categories: in a UMAP visualization, if two categories are separable in the projected space they will also be separable in the original space (Tunstall et al., 2022). We projected the 768-dimensional embeddings onto a 2-dimensional plane highlighting the differences in distributions for the four different classifier types.

The distributions of sortal and mensural classifiers are predicted to be alike if they occur with similar words around them, as suggested by several authors in existing literature (Lyons, 1977; Cheng and Sybesma, 1999; Paik and Bond, 2002; Bender and Siegel, 2004; Gebhardt, 2011, among others). Given the way the UMAP algorithm has been designed, if classifier distributions align, their UMAP projection should mostly overlap. Such an overlap would constitute an argument towards positing one single classifier category for all overlapping types. If classifier distributions do

Sortal:	**口**, **扇**, **盏**, **出**, **尊**, **棵**, **条**, **枝**, **所**, **张**, **粒**, **头**, **顶**, **把**, **面**, **封**, **管**, **道**, **件**, **匹**, **门**, **枚**, **堵**, **座**, **只**, **架**, **首**, **朵**, **家**, **篇**, **辆**, **卷**, **个**, **行**, **颗**, **杆**, **处**, **桩**, **幅**, **顿**, **部**, **幕**, **位**, **艘**, **根**, **本**, 株, 宗, 栋, 则, 支, 幢, 台, 项, 袭
Mensural: quantity	**级**, **片**, **盘**, **股**, **壶**, **脸**, **排**, **系列**, **册**, **手**, **号**, **层**, **团**, **段**, **周**, **版**, **包**, **瓶**, **帮**, **箱**, **堂**, **锅**, **节**, **岁**, **剂**, **组**, **块**, **种**, **轮**, **类**, **杯**, **天**, **盆**, **筐**, **盒**, **堆**, **桶**, **世**, **边**, **套**, **名**, **副**, **担**, **队**, **对**, **笔**, **页**, **派**, **味**, **划**, **群**, **截**, **袋**, **族**, **栏**, **脚**, **区**, **番**, **点**, **列**, **章**, **厘**, **票**, **分**, **路**, **班**, **些**, **站**, **批**, **月**, **丝**, **桌**, **阶**, **碗**, **年**, **重**, **肚子**, **双**, **身**, **代**, **盅**, **串**, **样**, **滴**, **缸**, **笼**, **辈**, **罐**, **眼**, 撮, 匙, 届, 垛, 竹篓, 尾, 筒, 篓, 坨, 集, 帧, 墩, 柄, 户, 扎, 刻, 餐, 具, 起, 发, 针, 品, 日, 小捆, 瓮, 屉, 酒杯, 杓, 句, 场, 炉, 竿, 樽, 沓, 簇, 期, 茶匙, 箩筐, 记, 席, 缸子, 间, 缕, 池, 阕, 囊, 员, 帖, 伙, 拨, 曲, 束, 圈, 辑, 叠, 波, 摊, 份, 楼, 款
Mensural: measurement	**秒**, **英里**, **米**, **克**, **西西**, **公尺**, **亩**, **公里**, **毛**, **丈**, **英寸**, **英尺**, **公斤**, **公分**, **小时**, **斗**, **码**, **海里**, **加仑**, **寸**, **吨**, **尺**, **磅**, **斤**, 平方米, 公吨, 呎, 平方英呎, 微米, 立方英尺, 兆瓦特, 毫克, 公克, 兆赫, 瓦, 兆, 度, 厘米, 平方公尺, 安培, 千伏, 平方公里, 元, 英吋, 盎司, 公升, 打, 立方米, 平方英尺, 盎斯, 平方尺, 海浬, 公厘, 克拉, 平方呎, 毫米, 毫升, 英哩, 千兆, 大卡, 千瓦时, 美分, 千伏特, 伏特, 英亩, 瓦特, 坪, 公顷, 摄氏度, 伏, 平方厘米, 海哩, 千克, 微秒, 浬, 兆瓦, 立方厘米, 平米, 砬, 吋, 平方海里, 公倾, 千瓦, 华里, 角, 兆瓦时
Mensural: currency	**先令**, **法郎**, **卢比**, 克朗, 英磅, 马币, 比索, 新币, 缅元,, 瑞典克朗, 银元, 加元, 铢, 丹麦克朗, 韩币, 台币, 澳元, 澳币, 港币, 日圆, 镑, 新元, 泰铢, 美金, 欧元, 美元, 港元, 日元, 英镑, 韩元, 澳大利亚元

Table 4: List of all 315 manually validated and classified classifiers mainly based on Chao (1965)' reference grammar. Classifiers explicitly mentioned in the grammar are indicated in bold face.

not align, the UMAP projections should present as largely distinct. This would favor an analysis of more than one existing syntactic classifier category.

2.2.2 Evaluating functional contributions

Given that classifiers precede nouns within noun phrases, we also wanted to test whether classifiers, like German gendered articles (Dye et al., 2017), contribute to reducing uncertainty about upcoming nouns, and whether this reduction is equally operated by all types classifiers previously identified. For that purpose, we applied the information-theoretic measure of mutual information (MI) (Cover and Thomas, 2012) to all classifiers and their corresponding head nouns.

Mutual information (MI) indicates how much information (in bits) is shared between a classifier and its corresponding head noun. The higher the value of mutual information for a specific classifier-noun pair, the more systematically those nouns can be found together. In terms of processing, this systematicity contributes to significantly reducing the listener's (or reader's) uncertainty about the upcoming noun. Low mutual information would on the contrary indicate that classifiers are not particularly helpful for predicting (a class of) upcoming nouns.

If C and N represent the sets of all classifiers and nouns respectively, and c and n their corresponding elements, then MI between each type of classifier and its corresponding nouns is defined as follows:

$$\begin{aligned} I(N;C) &= H(N) - H(N|C) \\ &= \sum_{n \in N, c \in C} p(n,c) log \frac{p(n,c)}{p(n)p(c)} \end{aligned} \quad (1)$$

We computed the mutual information between classifiers and nouns for each type of classifier. We then used a one-way ANOVA to evaluate the level of significance of the differences in MI across all four categories (sortal classifiers, quantities, measurements, and currencies).

3 Results

3.1 Distribution

The UMAP projections for the distributions of sortal classifiers and all three subtypes of mensural classifiers are plotted in Figure 1. Darker zones correspond to a higher proportion of projections in that area of the plane. Lighter zones correspond to fewer projections or an absence of projections in that area.

The plots show distinct patterns of distribution for all four different types of classifiers: while there may be some partial overlap in lighter regions, dark (high token frequency) regions are clearly separated out for all four subcategories. They all occupy different regions in the plane.

Currencies are especially well separated from the rest three subcategories. However, it is worth noting that the word embeddings for currencies are only a small subset of our full dataset since the majority of word embeddings for currencies are represented by more than one character corresponding to more than one vector in the BERT model. These multi-character tokens had been removed in the data pre-processing phase.

Both sortal classifiers and generic measure words (quantities) show a broader range of possible distributions. Yet the former are mainly clustered in the right region of the plane, while the latter are concentrated in the left half of the plane. The distribution of measurements is mostly situated in the middle.

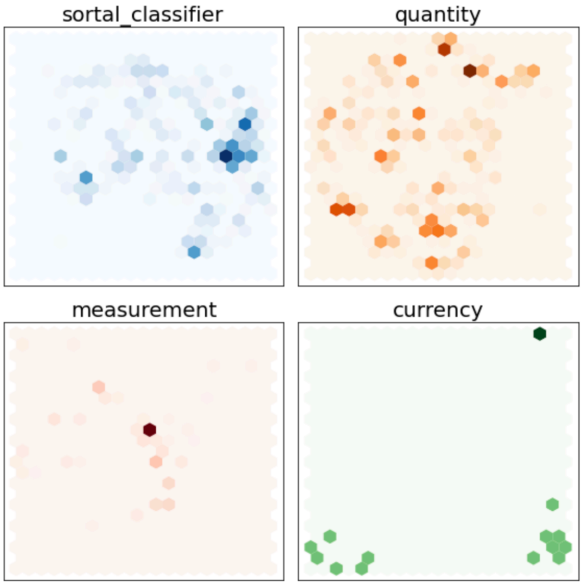

Figure 1: Visualization of the projections of 500,988 contextual word embeddings for all classifiers using UMAP (McInnes et al., 2018)

3.2 Function

For all classifier types, classifiers drastically reduce the entropy of upcoming nouns.

A one-way ANOVA test revealed that the difference in mean mutual information associated with each classifier type is significant across all four

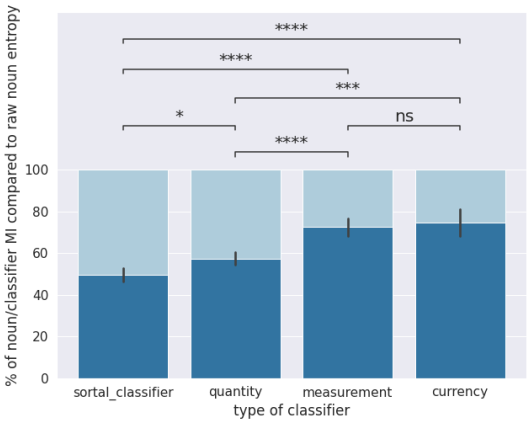

Figure 2: Percentages of Mutual information between nouns and classifiers over the entropy of noun. Error bars indicate bootstrapped (n sample = 10,000) *95% C.I.* of I(N; C). The number of asterisks denotes the magnitude of significance compared to a significance level of $p = 0.05$.

types.[7] We also used Tukey's HSD Test to perform multiple comparisons across the different types of classifiers.

We found that the mean mutual information between nouns and units of measurements is not significantly different from that between nouns and currency units.[8] Functionally, those two subtypes appear to play very similar roles: The overall entropy of nouns from our corpora lies around 12.29 bits. From Figure 2, it is apparent that both units of measurement (8.90 bits) and currencies (9.17 bits) greatly help with predicting upcoming nouns: knowing a measurement or currency accounts for around 75% of the original noun entropy. These two subcategories are significantly different from the other two.

MI involving units of measurements was significantly different from that involving sortal classifiers or generic quantities.[9] Not unsurprisingly, MI involving currency units also significantly differed from MI involving sortal classifiers or generic quantities.[10]

Even though the significance levels were not as high as for all other significant category pairs, differences in MI involving generic quantities vs.

[7] $F(3) = 25.46$, $P < 0.0001$.

[8] $P = 0.94$, *95% C.I.* = [-1.55, 1.02].

[9] $p = 0$, *95% C.I.* = [-3.80, -1.81]; $p = 0$, *95% C.I.* = [-2.63, -1.07].

[10] $p = 0$, *95% C.I.* = [-4.45, -1.70]; $p = 0.0001$, *95% C.I.* = [-3.35, -0.90].

those involving sortal classifiers still reached significance levels.[11] Our findings indicate that at the functional level, measure words can be distinguished from sortal classifiers. The presence of a measure word denoting generic quantities makes the upcoming noun more predictable than a sortal classifier in the same context. Classifiers denoting quantities (7.05 bits) account for 57% of the raw noun entropy, while sortal classifiers (6.09 bits) only account for 49%.

As a result, functional properties again suggest that mensural classifiers and sortal classifiers are better analyzed as two separate categories. Additionally, the results also suggest that the mensural classifier class is not homogeneous and that it may be better analyzed as at least two separate (sub-)categories: classifiers indicating generic quantities on the one hand and currency units and units of measurement one the other.

4 Discussion and relation to previous work

There is a longstanding debate as to whether mensural and sortal classifiers should be considered as the same grammatical category in Mandarin Chinese (or in classifier languages in general). Despite a general consensus that categorization should be performed on the basis of observable distributions and functions, researchers' conclusions diverge.

For some, sortal and mensural classifiers should be considered as one category since they can occur in similar contexts (e.g., Lyons, 1977; Cheng and Sybesma, 1999; Paik and Bond, 2002; Bender and Siegel, 2004; Gebhardt, 2011).[12] Others argue that sortal and mensural classifiers should be considered as distinct since they cannot be modified in the same way (Her and Hsieh, 2010; Singhapreecha, 2001; Nguyen, 2004). Her and Hsieh (2010) specifically argue that the difference between sortal and mensural classifiers is mainly semantic but has consequences on distributional properties: mensural classifiers are semantically substantive and block numeral quantification and adjective modification of the noun, whereas sortal classifiers are semantically null and not as restrictive.[13]

Similar differences are also claimed to exist in other classifier languages, such as Thai (Singhapreecha, 2001) or Vietnamese (Nguyen, 2004). In Vietnamese, for instance, mensural classifiers are described as sometimes occurring with modifiers inserted between the classifier and the head noun, whereas nothing can be inserted between a sortal classifier and its head noun. In general, previous literature arguing for separate categories for sortal and mensural classifiers tends to highlight that mensural classifiers can occur with more modifiers than their sortal counterparts. This is another way of saying that mensural classifiers are considered to allow a wider range of distributions than sortal classifiers.

Our results appear to corroborate that claim. Overall, in our results sortal and mensural classifiers do not appear to significantly overlap in their distributions, suggesting the existence of two distinct categories from a distributional perspective. But the UMAP plot in Figure 1 also shows more different distributions for mensural classifiers than for sortals, especially if generic quantities, currencies, and units of measure are analyzed as one group.

As an overarching category, mensural classifiers – including quantities, units of measurements, and currencies – appear to have a very diverse range of distributions. Given the very specific distributions for currencies and measurements, our data and the results of our analysis of classifier distributions appears to suggest that those might be better distributionally analyzed as three separate categories. Even if we only compare generic quantities to sortal classifiers, the range of projections associated with the mensural classifiers clearly exceeds that of the sortals, in line with conclusions drawn by proponents of separate syntactic categories.

From a functional perspective, some researchers have attempted to argue that mensural classifiers should be considered as belonging to the same syntactic category as their sortal counterparts because of the parallel roles they play within noun phrases (see Lyons, 1977; Cheng and Sybesma, 1999; Paik and Bond, 2002; Bender and Siegel, 2004; Gebhardt, 2011, among others).[14] The results of our

[11] $p = 0.03$, *95% C.I.* = [-1.86, -0.04].

[12] In Table 1, the sortal classifier 张 *zhāng* appears in a similar position as the mensural classifiers 组 *zǔ*, 斤 *jīn*, and 美元 *měiyuán* (between either a number or a determiner and a noun).

[13] Her and Hsieh suggest three diagnostic distributional tests to differentiate sortal and mensural classifiers: numeral/adjectival stacking modification, *de*-insertion, and *ge*-substitution.

[14] In Table 1 for example, both sortal classifier (张 *zhāng*) and mensural classifiers (组 *zǔ*, 斤 *jīn*, and 美元 *měiyuán*) can

study are closer in line with work suggesting that sortal and mensural classifiers are in fact functionally different.

Our study focuses on differences in the communicative function across classifier types. While we assume, based on evidence found for German gendered articles (Dye et al., 2017, 2018), that all classifiers will to some degree help anticipate the upcoming noun in the noun phrase they occur in, we wanted to test whether there would be a significant difference in the amount of MI effectively shared between the classifiers and their head noun. Such a significant difference would then suggest the existence of multiple syntactic categories associated with classifiers.

Related work by Liu et al. (2019) used MI to investigate how systematically classifiers can be predicted from the semantics of a given noun. The answer to that question would be relevant to questions related to classifier learning, but is distinct from our study. By focusing on the relation between noun entropy and its reduction in the presence of a classifier, we are specifically targeting the predictive value of classifiers in noun phrase processing.[15]

Our results show that there are significant differences in how much different types of classifiers help predict upcoming head nouns, with currency and measure units being the most predictive, classifiers denoting generic quantities ranking second, and sortal classifiers being the least helpful. Interestingly, while our results do suggest the existence of three different classifier categories from a functional perspective, the observed functional contributions are the opposite of what previous literature would have suggested.

Proponents of distinct classifier categories typically argue that while sortal classifiers are associated with nouns based on their referents' inherent properties (such as shape, humanness, animacy, etc.), mensural classifiers denote quantities not directly related to the nouns' meanings (see for example Jarkey and Komatsu, 2019; Unterbeck, 1994), suggesting that sortal classifiers would be more specifically linked to the nouns they combine with.[16] What we see in the results of our MI calculations is that all mensural classifier types share a greater amount of information with their head nouns than sortal classifiers do.

5 Conclusion

The distinction between sortal and mensural classifiers has been a long-standing debate in the fields of Chinese, (South-)East Asian linguistics, general linguistics and linguistic typology. Previous literature attempted to solve this problem using isolated example sentences and categorical grammaticality judgements. In this paper, we instead systematically re-evaluate the distributional and functional properties of classifiers using quantitative methodologies.

Using 981,076 noun phrases from a 489MB dependency-parsed corpus of Mandarin Chinese, we show that mensural and sortal classifiers are indeed measurably different both in their distributions and their functional contribution to noun phrase processing. We further find that mensural classifiers do not constitute a homogeneous class. Based on both their very specific distributions and they very significantly different functional contributions, units of measurement and currency can be classified as one if not two classes that are distinct from both sortal classifiers and generic measure words.

Our results also include two broader typological implications: since (i) sortal and mensural classifiers can be reliably identified as distinct categories in at least one language, (ii) the most promising line of analysis for further typological investigations into classifier systems will investigate whether languages with classifier systems cluster into two discrete types: those with separate categories for sortal and mensural classifiers, and those without a clear sortal/mensural split.

6 Appendix

Limitations and future work

In our results, currencies appeared as distributionally very different from both other mensural classifier types. However, when we extracted the contextual word embeddings of the classifiers for the distributional analysis, we discarded word embeddings for multi-token classifiers since they would

be used to quantify nouns.

[15] Lau and Grüter (2015) also investigate classifiers from a processing perspective, but using an experimental approached based on eye-tracking experiments involving L2 speakers of Mandarin.

[16] E.g., in table 1 the sortal classifier 张 *zhāng* combines with the referent/noun 地图 *dìtú* 'map' highlighting its flat properties, while the mensural classifier 斤 *jīn* quantifies the referent/noun 米 *mǐ* 'rice' by applying a specific measuring unit.

be represented by multiple rather than a single vector. This significantly reduced the number of representations for currencies. In the future, we might be able to use average vectors over multi-tokens or leftmost vectors to represent those discarded currencies, but further work will be needed to show their specific distributions. Regardless of this limitation, our study still revealed significant differences between the two largest subsets of classifiers: sortal classifiers and generic measures of quantity.

Our data does not cover all possible types of written and spoken genres. Yet, since a limited sample of genres already reveals distributional and functional differences between the two types of classifiers, those differences justify assigning sortal and mensural classifiers to separate syntactic categories in Mandarin Chinese. Future work could compare results across a broader variety of genres, notably to investigate classifier use in spoken Mandarin Chinese, where speakers may be more likely to either drop classifiers or make more extensive use of the most common generic classifier 个 *gè* at the expense of all other classifiers.

This project focuses on classifiers in Mandarin Chinese. In the future, we may be able to apply this methodology to other classifier languages to assess whether split classifier systems are the norm for languages with classifier systems or whether languages cluster into two discrete types: those with separate categories for sortal and mensural classifiers, and those without a clear sortal/mensural split.

Our code will be made available for replication and extension by the community.

Acknowledgements

This project was supported by resources provided by the Office of Research Computing at George Mason University (URL: https://orc.gmu.edu) and funded in part by grants from the National Science Foundation (Awards Number 1625039 and 2018631, and 2109578).

References

Alexandra Y Aikhenvald and Elena I Mihas. 2019. *Genders and classifiers: a cross-linguistic typology*. Oxford University Press.

Emily M Bender and Melanie Siegel. 2004. Implementing the syntax of japanese numeral classifiers. In *International Conference on Natural Language Processing*, pages 626–635. Springer.

Yuen Ren Chao. 1965. *A grammar of spoken Chinese.* ERIC.

Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750.

Lisa Lai-Shen Cheng and Rint Sybesma. 1999. Bare and not-so-bare nouns and the structure of np. *Linguistic inquiry*, 30(4):509–542.

Thomas M Cover and Joy A Thomas. 2012. Elements of information theory. 2012. *Google Scholar Google Scholar Digital Library Digital Library*.

Matthew Dryer, David Gil, and Martin Haspelmath. 2005. *The world atlas of language structures*. Oxford University Press.

Melody Dye, Petar Milin, Richard Futrell, and Michael Ramscar. 2017. A functional theory of gender paradigms. In *Perspectives on morphological organization*, pages 212–239. Brill.

Melody Dye, Petar Milin, Richard Futrell, and Michael Ramscar. 2018. Alternative solutions to a language design problem: The role of adjectives and gender marking in efficient communication. *Topics in cognitive science*, 10(1):209–224.

Lewis Gebhardt. 2011. Classifiers are functional. *Linguistic Inquiry*, 42(1):125–130.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765.

One-Soon Her and Chen-Tien Hsieh. 2010. On the semantic distinction between classifiers and measure words in chinese. *Language and linguistics*, 11(3):527–551.

Nerida Jarkey and Hiroko Komatsu. 2019. Numeral classifiers in japanese. *Genders and classifiers: a cross-linguistic typology*, pages 249–81.

Elaine Lau and Theres Grüter. 2015. Real-time processing of classifier information by l2 speakers of chinese. In *Proceedings of the 39th annual Boston University conference on language development*, pages 311–323.

Charles N Li and Sandra A Thompson. 1989. *Mandarin Chinese: A functional reference grammar*, volume 3. Univ of California Press.

XuPing Li. 2013. Numeral classifiers in chinese. In *Numeral Classifiers in Chinese*. De Gruyter Mouton.

Shijia Liu, Hongyuan Mei, Adina Williams, and Ryan Cotterell. 2019. On the idiosyncrasies of the mandarin chinese classifier system. *arXiv preprint arXiv:1902.10193*.

John Lyons. 1977. *Semantics: Volume 2*, volume 2. Cambridge university press.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Tuong Hung Nguyen. 2004. *The structure of the Vietnamese noun phrase*. Boston University.

Kyounghee Paik and Francis Bond. 2002. Spatial representation and shape classifiers in japanese and korean. *The Construction of Meaning*, pages 163–180.

Pornsiri Singhapreecha. 2001. Thai classifiers and the structure of complex thai nominals. In *Proceedings of the 15th Pacific Asia Conference on Language, Information and Computation*, pages 259–270.

Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2022. *Natural language processing with transformers*. " O'Reilly Media, Inc.".

Barbara Unterbeck. 1994. Korean classifiers. *Theoretical issues in Korean linguistics*, pages 367–385.