George STAROSTIN
Higher School of Economics / Russian State University for the Humanities / Santa Fe Institute; gstarst1@gmail.com

Chinese basic lexicon from a diachronic perspective: implications for lexicostatistics and glottochronology[1]

In this paper, I attempt to compare the relative rates of replacement of basic vocabulary items (from the 100-item Swadesh list) over four specific checkpoints in the history of the Chinese language: Early Old Chinese (as represented by documents such as *The Book of Songs*), Classic Old Chinese, Late Middle Chinese (represented by the language of *The Record of Linji*), and Modern Chinese. After a concise explication of the applied methodology and a detailed presentation of the data, it is shown that the average rates of replacement between each of these checkpoints do not significantly deviate from each other and are generally compatible with the classic «Swadesh constant» of 0.14 loss per millennium; furthermore, these results correlate with other similar observed situations, e.g. for the Greek language, though not with others (Icelandic). It is hoped that future similar studies on the lexical evolution of languages with attested written histories will allow to place these observations into a more significant context.

Keywords: Chinese language history, Old Chinese language, Middle Chinese language, lexicostatistics, glottochronology, basic vocabulary.

Introduction

Over the last couple of decades, lexicostatistical methodology has played an important role in historical studies on the evolution of various «dialectal» forms of Chinese (or «Sinitic languages», from a more strictly linguistic point of view). Since there is no universally accepted model of the lexicostatistical procedure as far as the selection of source data, manual and/or automated annotation of lexical cognates, and the specific phylogenetic algorithm applied to the data are concerned, these studies significantly vary in terms of selected scope, stated goals, and attained results; but there seems to be a general understanding that conducting lexicostatistical studies is an important stage in unraveling the internal history of Chinese and identifying certain key points resulting in divergent linguistic lineages, as well as separating evidence for genetic splits from evidence for later linguistic contacts that tend to obscure the different lineages in question.

That said, most of the studies on Chinese (Sinitic) lexicostatistics have largely focused on quantifying and interpreting the degree of lexical divergence between modern colloquial forms of Chinese[2], usually downplaying the important fact that Chinese is one of the very few

[1] I thank Prof. Laurent Sagart for his valuable comments on parts of this paper, and Dr. Johann-Mattis List for the opportunity to present its major points before a large audience of specialists at the *Old Chinese And Friends* conference (Max Planck Institute for the Science of Human history, Jena, April 26–27, 2018). Any errors in data or its analysis are exclusively my own.

[2] It is not within the scope of the current paper to provide a detailed listing of all the works that have applied quantitative methods to the problem of Chinese dialect classification. For those unfamiliar with the topic a good starting point could be the complex study of Mahé Ben Hamed and Wang Feng (2006), who apply a variety of dis-

languages in the world whose historical evolution can actually be studied by means of preserved written data, rather than reconstructed through the comparative method — and, consequently, one of the most important test cases in the world (along with several Indo-European and Semitic languages) when it comes to measuring rates of lexical evolution[3].

The reasons for such negligence are understandable. Studying lexical replacement in languages represented only by a closed and limited corpus of written data necessarily runs into a number of uncertainties — insufficient attestation of required items in available texts, their occasional semantic ambiguity, and lack of direct knowledge on the dialectal characteristics of said texts, among other things. To make matters worse, historically attested forms of Chinese are commonly understood to mix together different strata — to the point that, for instance, our current understanding of Middle Chinese phonology (as extracted from rhyme books and rhyme tables) vastly exceeds our understanding of Middle Chinese grammar *and* lexicon, since most texts in the classic era of Táng and early Sòng dynasties were written in one or another variant of the archaic Literary Chinese. Circumstances such as these may seem to make the painstaking task of studying lexical replacement within Chinese in detail a waste of time, but in reality it is not that difficult to employ a somewhat formalistic approach to the matter and at least try to see what it gets us. However, in order for such a study to be of any use, it is imperative to state the rules very clearly and consistently apply them to all selected time periods and data collections.

The present paper is a tentative attempt to manually measure the rates of lexical evolution over a period of approximately 2,500–2,800 years in the history of Chinese. This is achieved by selecting several chronological checkpoints, constructing standardized Swadesh wordlists for each of them and individually investigating each certified or potential case of lexical replacement from one checkpoint to another. Two reasons why such a study, though still clearly far from perfect, could not have appeared earlier, are as follows: (a) a breakthrough in corpus studies on Old Chinese — largely due to the outstanding dedicated work of Donald Sturgeon and his colleagues, we now have the advantage of the online *Chinese Text Project*, allowing for complex lexical investigations on a large scale to be conducted almost momentarily; (b) significant methodological clarification of the lexicostatistical technique, described in several papers from the Moscow school of comparative linguistics (see the "Methodology" section below). Naturally, there is still much room for improvement (especially in the area of Middle Chinese, which remains considerably underdeveloped), but there is reason to believe that even at this stage, the results will be useful enough both for Sinologists and general specialists in diachronic linguistics.

Before presenting the data in its entirety, it is necessary to do the following things: (a) justify and describe the four selected chronological checkpoints — Early Old Chinese, Classical

tance- and character-based methods in order to determine whether the configuration of known forms of Chinese better agrees with a tree-like or a network-like structure; the same data was later made use of by Johann-Mattis List (2015) in his own investigation of the historical relations between Chinese dialects. Further references to earlier studies may be easily found in either of those papers.

[3] To the best of my knowledge, only two brief attempts at measuring the lexical distance between Old Chinese and Modern Chinese have had their results mentioned in literature: (a) Swadesh 1952: 456 and subsequent papers by both Swadesh and Robert Lees make frequent reference to the results of C. Y. Fang, who allegedly established that 79% out of the 200-item wordlist of «Classic Chinese 950 A.D.» have been retained in «modern colloquial Northern Chinese»; the wordlist itself has never been published, making it impossible to verify the claim, and it is in fact quite unclear what is meant by «Classic Chinese 950 A.D.»; (b) Starostin 2000: 256 actually gives a specific list of replacements from «Archaic Chinese (seventh century BC)» to modern Mandarin that can be checked, and the verification shows a significant number of omissions (see below for specific examples).

Old Chinese, (Late) Middle Chinese, and Modern Chinese, including some discussion on dating, data sources, and various technical problems; (b) give a brief description of the methodology employed in selecting items for the respective positions in the wordlist, as well as the procedure of cognate scoring from one period to another. This will be followed by reasonably detailed discussion of the data itself, after which we present a brief analysis and state our conclusions on the tendencies of lexical evolution in the history of Chinese, including a typological-comparative angle.

Data sources

1. Early Old Chinese (EOC)

Definition: we approximately define Early Old Chinese as the language that is represented in writing by such literary monuments as the *Shījīng* ('Book of Odes') and the oldest parts of the *Shūjīng*, or *Shàngshū* ('Book of Documents'), as well as epigraphic data from artefacts (mainly bronze vessels) dating back to the Early Zhōu dynasty (*jīnwén*); the most comprehensive and systematic Western dictionary of this language is Schuessler 1987. In general, the language of all these texts is known to share certain grammatical and lexical properties that strongly distinguish it from later forms of Chinese, though it cannot be said for certain to represent a direct ancestral stage for any of them.

Reasons for selection: EOC is the very first chronological checkpoint for which it is possible to construct anything close to a standardized Swadesh wordlist. Although some observations may be made on certain elements of the basic lexicon in the oracle bone inscriptions of the Shāng dynasty, the restricted and highly formulaic nature of these inscriptions leads to way too many gaps in the wordlist for it to be of any use for the present study. Therefore, a general statistically relevant investigation of Chinese basic lexicon may only begin from Early Zhōu times.

Sources: Much, if not most, of the epigraphic material is ineligible for the task of building a Swadesh wordlist due (once again) to the highly formulaic subject matter and ritualistic nature of the texts, leaving the verses of the *Shījīng* as the single most natural source for an EOC list of basic lexicon. Out of the 100 required items, only eight ('ashes', 'bark', 'bone', 'egg', 'knee', 'lie', 'liver', 'louse') have no reliable or probable equivalents attested in the *Shījīng* (or in the eldest parts of the *Shūjīng*).

Problems: There is little doubt that the texts of the *Shījīng* are relatively heterogeneous in terms of both time and space (see Dobson 1968: 224–242 for an attempt at a chronological linguistic stratification of the various sections of the *Shījīng* based on grammatical evidence), but there is so far very little evidence that the dialects of the *Shījīng* were significantly different from each other as far as their basic lexicon was concerned: very few synonyms for basic notions were elicited from the data, and those that were elicited are not easily described in terms of dialectal variety (see, e.g., 'give' below, with two different synonyms attested in the exact same poem). A much more significant problem is the scarceness of attestation for multiple terms: in many cases unambiguous contexts with the required word are found but once or twice, and their reliability often depends on external data (e.g. if the same word is also the basic equivalent for the term in Classical Chinese, this improves the chances of the corresponding item in EOC). All such terms are specially commented upon in the notes section, and particularly dubious inclusions are marked with a question sign.

2. Classical Old Chinese (COC)

Definition: We define COC as the language of literary texts, most likely reasonably close to the spoken language of the time, written from approximately the end of the 5th century to the end of the 3rd century BC. There is no single defining dictionary for this stage of the language, since lexicographical sources usually conflate it either with EOC or with Hàn-era OC (or both); however, the text corpus is reasonably well defined, and focused searches may be performed these days with the aid of such resources as the *Chinese Text Project* (Sturgeon 2019).

Reasons for selection: COC is the first known historical stage of Chinese that is represented by a substantial amount of thematically diverse non-poetic texts which, according to a general scholarly consensus, are written in a language that reasonably closely reflects colloquial patterns of the time (with certain expected stylistic emendations, though their influence on core basic lexicon is probably negligible). A significant advantage of this period is that the language of the texts in question is not as highly influenced by the language of the previous period (EOC) as the written language of Hàn-era and later periods is dependent on COC.

Sources: COC is generally understood to have possessed a significant amount of dialectal diversity; even if evidence for this rarely comes from core basic vocabulary, for the sake of increased accuracy we prefer to draw upon sources typically recognized to stem from the same dialectal area. The principal texts corroborating our selections are the *Lùn yǔ* and (especially) *Mèng-zǐ*, both recognized as representative of the Lǔ dialect (Pulleyblank 1995: 3), although there may be a chronological gap of about 100–150 years in their original composition (not essential for our purposes).

If the necessary words are encountered very rarely or not encountered at all in these texts, we find it acceptable to draw upon data from other sources, such as the *Zuǒ zhuàn* (representing a separate dialect of its own, together with the *Guó yǔ*) and *Zhuāng-zǐ* (probably representing a more Southern, Chǔ-area, dialect, though this is debatable). For the record, the following words are not attested at all in the *Lùn yǔ* and *Mèng-zǐ* and have to be substituted from other sources: 'ashes', 'bite', 'nail', 'dry', 'green', 'knee', 'liver', 'louse', 'red', 'root', 'round', 'sand', 'smoke', 'swim', 'tail', 'tongue'. Since every single one of these 16 items is either the same as in EOC or the same as in MC or both, and since we have been unable to reliably elicit even a single undeniable difference in the Swadesh wordlist between any of the listed texts, such substitution should be permissible.

Problems: COC is (arguably) one of the least problematic periods in the history of Chinese when it comes to eliciting basic lexicon; see above on the relative insignificance of dialectal divisions for this purpose. Several dubious cases of elicitation, usually having to do with scarceness of attestation and ambiguity of translation, are commented upon specifically in the data section of the paper.

3. Middle Chinese (MC)

Definition: For the purposes of the current study, Middle Chinese is narrowly defined as the colloquial (or reasonably close to colloquial) stage of Chinese, chronologically coinciding with or closely following the beginning of the division of Chinese into the principal dialectal groups of today, i.e. what is commonly called *Late* rather than *Early* Middle Chinese. This is due to the fact that texts from the Early Middle Chinese era (first half of the Táng dynasty, 7th–8th centuries AD) are nearly always written in an archaic form of the language (*wén yán* or *gǔ wén*), whereas for the Late Middle Chinese period (late Táng and early Sòng dynasties) there is a limited, but useful corpus of textual evidence that is somewhat sufficient for purposes of lexicostatistical analysis.

Reasons for selection: The entire period between COC and the 20th century is an extremely difficult area for lexicostatistical evaluation, since almost every text written in traditional imperial China, from Hàn all the way to Qíng dynasties, is influenced, to various degrees, by the grammar and lexicon of COC, and hardly ever reflects the spoken language of the corresponding period. It is precisely for this reason that we have refrained, for instance, from attempting to construct a separate 100-item wordlist for the language of the early or late Hàn dynasty, despite the abundance of textual evidence from that period — perusal of such vast sources as Sīmǎ Qiān's *Shǐ jì*, for instance, shows that in many cases Swadesh items are represented by at least two competing equivalents (e.g. 犬 *quǎn* and 狗 *gǒu* for 'dog', 盈 *yíng* and 滿 *mǎn* for 'full', etc.), and it is often impossible to determine whether such situations are due to true «transit synonymy» (when a lexical innovation has not yet fully managed to displace the original neutral term) or to the intentional (or unintentional) mixing of standard colloquial and outdated archaic equivalents.

Nevertheless, it is very important to have at least one analyzable «checkpoint» on the almost 2,500 year long way from COC to Modern Chinese, and from a general chronological and qualitative point of view, Late Middle Chinese is the optimal, if far from perfect, candidate for this purpose, since this is the period of proliferation for the genre of the 語錄 *yǔlù* («records of sayings»), a new genre of Buddhist literature whose innovative and frequently iconoclastic nature placed a large emphasis on transmitting sermons, parables, and real life anecdotes by means of colloquial idioms. In general, the *yǔlù* may be considered as the first fully colloquial genre of literature in the history of past-COC Chinese, and although it is more thematically limited than the fictional genres of late Sòng, Yuán, and Míng dynasties, its advantages are that it is represented by chronologically older texts and that at least some of these texts are arguably more free from literary embellishments than the literary genres of *huàběn* and *xiǎoshuō* (classic short stories and novels from Sòng to Míng-Qíng times).

Sources: A thorough lexical analysis of all or most of the existing texts in the *yǔlù* genre has not been conducted yet; an important problem is that some of the texts may reflect serious dialectal differentiations. For this reason, analysis has so far been restricted to just one reasonably large and generally uniform specimen of the genre, namely, the *Línjì yǔlù* («The record of Linji»), a text traditionally attributed to the disciples of the school of Master Linji Yixuan (d. 866 AD) but not finalized until the late 11th–early 12th centuries. The language of the *Línjì yǔlù* and the *yǔlù* genre has been the subject of several meticulous studies, e.g. Sawer 1969, Gurevich 2001, but all of them focus almost exclusively on grammar rather than lexicon; nevertheless, analysis of the basic words used in the text is in perfect agreement with the grammatical data in that the *Lǐnjì yǔlù* does indeed attempt to represent the colloquial standards of its time, albeit with some inescapable influence of the more classical forms as well.

Problems: Restriction to a single source necessarily implies that our MC list has the heaviest gaps of all (at least 18 out of 100 items are not featured at all in the text, and 8 more are somewhat problematic due to scarceness of attestation and semantic ambiguity); the problem is somewhat alleviated by the fact that the majority of these gaps are items that are represented by the same equivalent in COC and Modern Chinese, so it may be reasonably assumed that they were not replaced by anything else in MC as well.

4. Modern Chinese (PTH)

Definition: Since this study is only concerned with the issue of relatively straightforward diachronic evolution from a single point in the past to a single point in the present, we intentionally limit our definition of «Modern Chinese» to the present day version of *pǔtōnghuà*, the

common national language generally based on the Běijīng Mandarin dialect; linguistic differences between the actual spoken varieties of Běijīng Mandarin and *pǔtōnghuà* are well known, but do not generally extend to core basic vocabulary, making this factor negligible.

Reasons for selection: Theoretically, any other Chinese «dialect» (with the exception of Mǐn, since that cluster is typically assumed to have split off from the rest before the beginning of the MC period) might have been substituted here, but the task of constructing a 100-item wordlist for *pǔtōnghuà* is naturally easier than for any of the rest. A separate study is necessary to assess the rate of evolution from MC to PTH relative to other varieties of spoken Chinese that are in use today.

Sources: A variety of sources has been used (textbooks, dictionaries, text corpora, live informants etc.).

Problems: This is the least problematic area of all; issues are typically limited to purely technical problems, such as choosing a monosyllabic or bisyllabic variant for the most common equivalent of a given meaning (where the adopted solution usually bears no impact on calculations anyway).

Methodology of wordlist construction and lexical comparison

In constructing the optimal wordlists for each of the four stages, I attempt to follow as closely as possible the guidelines laid down in Starostin 2010 and Kassian et al. 2010, which can largely be boiled down to the following principles: (a) elicit words whose meaning and stylistic register are as close as possible to the pre-defined meanings listed in the latter paper; (b) try to avoid the inclusion of multiple synonyms, whose presence undermines the main idea of lexicostatistics.

Obviously, when dealing with written stages of the language represented by closed (and usually not very large) corpora, formal and precise adherence to these principles is not always possible. Due to the nature of the data itself, all of the wordlists presented below, with the exception of the wordlist for Modern Chinese, will inevitably contain errors, some of which might not even be rectified in the future unless massive new amounts of data (e.g. from archaeological sources) become available. However, the important thing here is to make certain that these errors do not skew the quantitative conclusions in any one particular direction, i.e. that they do not increase specifically the number of lexical replacements or the number of lexical retentions from any chosen point in the history of Chinese to the next one. This implies the necessity of a transparent, objective, well-argued methodology of dealing with ambiguous situations, one that should preferably minimize the possible interference of the personal preferences of the compiler. Below I list some of the general points; specific applications may be found in the comments on particularly troublesome lexical items in their respective sections.

1) *Be wary of etymological arguments*. Frequently, when facing the choice between picking one out of two or more synonyms, or including all of them into the list, one may be led astray by the fact that an older equivalent, clearly going back to the original main equivalent for a given Swadesh term, is still preserved at a later stage in the development of the language — ignoring the clear fact that its semantics has shifted, as the word is now used in a slightly different meaning or has been relocated to a different (marked) stylistic register (vulgar or elevated).

This is, for instance, the reason of several important mistakes in Starostin 2000: 256, a general study in the methodology of lexicostatistics where Old Chinese is compared with Modern Chinese and 23 lexical replacements are identified. The study fails to list several transparent

replacements, such as 目 *mù* → 眼睛 *yǎn-jing* 'eye' and 首 *shǒu* → 頭 *tóu* 'head', presumably because the former equivalents are still encountered today in various bound idiomatic formations and archaic contexts. This leads to underestimations of the process of lexical replacement, and the problem gets even worse for eras that are only represented by written documents, since written language by its very nature fails to keep up the pace with developments in the colloquial idiom, and special care must be given to the study of preserved texts in order to make a qualified decision on whether a certain lexical replacement has already been completed at a given period or not. In any case, 'etymological argument' alone, not supported by actual data from texts, does not carry significant value.

2) *Watch out for bound forms and idiomatization*. The «basic» equivalent of any given meaning is typically understood as the most neutral and generally context-independent form: the more words there are that an observed candidate can enter in syntactic relations with, the better are its chances for historical stability. Thus, COC has multiple equivalents for the meanings 'die' and 'kill', but a great majority of them has limited syntactic applicability: e.g. 弑 *shì* 'to kill' is only used in reference to killing a superior (prince, father, etc.), whereas 薨 *hōng* 'to die' is only said of high officials. Not surprisingly, these are precisely the words that do not survive into the MC era, whereas the neutral 殺 *shā* 'kill' and 死 *sǐ* 'die' persist all the way into most modern forms of Chinese.

3) *Textual evidence is generally superior to dictionary information*. With a closed and relatively limited textual corpus that is not particularly well reflected in specialized dictionaries, OC is clearly one of those ancient languages where direct elicitation of lexical data from the corpus is much preferable to relying on dictionaries. In a few cases, observations of actual word usage in the attested texts may lead to startlingly unpredictable conclusions (see notes on possible replacements from EOC to COC below); more importantly, finding relevant syntactic and semantic contexts adds a much wanted level of confidence to our wordlists, and also helps differentiate between statistically frequent and rarely used synonyms. This is particularly helpful for transitional stages of the language, in which an older equivalent may already be retained only as a rare archaism (including quotations from and paraphrasing of older texts), while the newer replacement may be more frequent — however, such situations will rarely, if ever, be discussed or even hinted at in dictionaries.

Regarding the procedure of cognate scoring, in this particular setting it is essentially reduced to the procedure of *postulating lexical replacements from one time period to another*. In addition to the obvious (lexical replacements are assumed whenever word X, used in a given Swadesh meaning over the time period t_n, is no longer used in that meaning over the time period t_{n+1}), we try to observe the following rules:

1) *Statistics and stylistics matter*. This is essentially a recapitulation of points 1 and 2 from the previous section: even if the same word is encountered seemingly in the same meaning over several distinct time periods, this does not always imply that it has not actually been replaced. Written Chinese has always operated according to the «forget nothing» principle: no matter how archaic a certain word is, there is always some probability of encountering it in texts that are separated by any number of years from its time of proliferation. What matters is primarily the statistics of usage (if there are two or more synonyms, which one is the most frequent?) and the stylistic context of usage (if there are two or more synonyms, which ones are used in quotations, poetic formulas, imitations of archaic rhetorics — and which ones are used in colloquial direct speech or neutral descriptions of situations?). If it can be shown that synonyms A and B express the same meaning in t_{n+1} as exclusively A in t_n, but that A is rare compared with B *and* primarily used in stylistically marked contexts, we postulate a clear-cut lexical replacement.

2) *Morphological change does not matter*. The issue of «partial cognacy», where two equivalents of the same Swadesh meaning in two different languages (or different stages of the same language) consist of two or more morphemes, of which only one (usually the root) is etymologically shared between them, while the others are different, seems to be particularly acute for languages that frequently resort to compounding techniques, including Chinese. This issue has been discussed several times in literature (e.g. List 2016; Starostin 2013a), but still remains without a perfect solution. Should a difference such as COC 知 *zhī* 'to know' vs. Modern Chinese 知道 *zhī-dào* id. be reflected by assigning both items the same index of cognacy (no lexical replacement), different indexes (replacement), or marked in some other manner (e.g.awarded «half a point» instead of a regular full +1 index, etc.)?

In my opinion, a definitive solution to this issue is impossible until a solid experimental base for this type of situations has been built up — which would allow us to cross-linguistically compare replacement rates for different methods of scoring and choose the solution that would make more general sense from a historical point of view. In the meantime, for Chinese I prefer to stick to the «no lexical replacement for partial cognacy situations» scenario, for the following reason: in most cases, morphemic compounding in the history of Chinese is explainable by reasons that have nothing to do with semantic shifts and more to do with the phonetic evolution of the language (avoidance of ever-increasing levels of homonymy), which is clearly not what we really want to measure when choosing lexical change as a base parameter for glottochronology. Therefore, in this study classical 知 *zhī* will be scored exactly the same as modern 知道 *zhī-dào*.

However, one important thing about both classical and modern Chinese compounds («binomes») that should be stated is that in many (not all) cases a binome may easily be analyzed as containing a *primary* and a *secondary* morpheme. The *primary* morpheme is the historical root morpheme; its defining diachronic characteristic is that it tends to be more stable over both time and space, and its defining synchronic characteristic is that, unlike the secondary morpheme, it can still be frequently encountered, usually in bound form, without the secondary morpheme in its original meaning. The *secondary* morpheme largely acts as an additional determiner: as a rule, it is less stable across time periods and dialects, it may be omitted in certain contexts, and whenever encountered on its own, it is rarely or never used in the same meaning as the primary morpheme.

A good example is Modern Chinese 月亮 *yuè-liàng* 'moon', where 月 *yuè* is the primary morpheme because it may be encountered on its own in the same meaning (usually in other bound forms, e.g. 月夜 *yuè-yè* 'moonlit night, etc.), whereas 亮 *liàng* 'light, shine' is never encountered with the meaning 'moon' if not in conjunction with 月 *yuè*; not surprisingly, 月 *yuè* is also the historically stable morpheme 'moon', common for most varieties of Chinese, whereas 亮 *liàng* is a more recent addition and alternates with other additions in different dialects (e.g. 月光 *yuè-guāng*, 月子 *yuè-zi* etc.).

Somewhat more complicated are cases of concatenated binomes in which, upon first sight, both morphemes express the same meaning and are hard to classify as respectively primary and secondary — such as 道路 *dào-lù* 'road' (literally 'road$_1$' + 'road$_2$') or 牙齒 *yá-chǐ* 'teeth' ('tooth$_1$' + 'tooth$_2$'). It would seem that technically, the best solution here would be to judge the two morphemes as synonymous and include both into the calculations. However, even in this situation analysis of the behavior of the respective meaning in different contexts actually shows that one morpheme typically prevails over the other. Thus, in the meaning 'road' Modern Chinese frequently employs simple 路 *lù* (大路 *dà lù* 'big road', etc.), but practically never 道 *dào* (which is far more common in the abstract meaning 'way, manner'); the meaning 'tooth / teeth' is frequently expressed by 牙 *yá* (as in 刷牙 *shuā yá* 'brush one's teeth', etc.), but almost

never by 齒 *chǐ*. I interpret this as clear evidence that in forms such as 道路 *dào-lù* and 牙齒 *yá-chǐ*, one morpheme still behaves as primary and the other one as secondary, even if from a historical point of view (as can be seen from comparison with OC evidence, see the data below) it is the secondary morpheme that reflects the original Swadesh equivalent — see, however, the «be wary of etymological arguments» point above, which clearly pressures us into regarding such situations as lexical replacements.

One might argue that such a solution directly contradicts the «morphological change does not matter», but this is only if we understand the dynamic genesis of such compounds as 牙齒 *yá-chǐ* as the extension of the primary morpheme 齒 *chǐ* with the «prefixed» quasi-synonymous morpheme 牙 *yá*, when in reality the process must have been far more complex: equivalents of the monosyllabic 牙 *yá* are found in the basic meaning 'tooth' in many Chinese dialects, as well as alternate binomes such as 牙巴 *yá-ba*, etc., indicating that the structure of 牙齒 *yá-chǐ* is, in fact, quite analogous to that of 月亮 *yuè-liàng*. Ignoring this would mean ignoring an important element of lexical restructuring in the history of Chinese, and while other formal solutions are possible, in this study we will try to consistently apply this principle to the procedure of cognate scoring.

Notes on transcription

Since this study is only concerned with different stages of Chinese and not with the Sino-Tibetan (or areal) origins of the Chinese entries, issues of phonetic and phonological reconstruction of OC and MC are largely irrelevant; cognate identification is not required between OC, MC, and PTH, and phonological or phonetic transcriptions of Chinese characters only matter inasmuch as the paper might also interest general historical linguists with no knowledge of Chinese hieroglyphics, or, occasionally, to specify which particular pronunciation out of several possible ones is meant for a specific character (e.g. 長 **draŋ* > *cháng* 'long', not 長 **traŋʔ* > *zhǎng* 'grown-up', etc.).

Throughout the study, I consistently use the OC reconstruction of Sergei Starostin (1989), some of the aspects of which remain controversial (e.g. the reconstruction of lateral affricates and voiced aspirates, or the interpretation of Type A / Type B syllable distinction as reflecting an opposition in vowel length) but which I also find reasonably conservative in comparison with the far-reaching changes in Baxter, Sagart 2014. OC Reconstructions are taken either directly from Starostin 1989 or from Sergei Starostin's unfinished etymological database on Old Chinese («Chinese Characters Database» at the Tower of Babel website, http://starling.rinet.ru). MC readings are used very sparsely throughout the rest of the paper; where necessary, they are also taken from Starostin's database. Modern Chinese forms are transcribed in standard *pinyin*. OC and modern readings are typically given back-to-back next to the respective characters, with OC reconstructions accompanied by asterisks.

The data

All four wordlists have been published online as part of the Sinitic 100-item wordlist database, included in the Global Lexicostatistical Database framework (http://starling.rinet.ru/new100); in addition to the words themselves, the database includes plenty of annotations and comments, such as precise references to sources, quotations of contexts from which the items have been elicited, and (sometimes highly detailed) explanations on why certain synonyms were pre-

ferred over others. This section of the paper represents a seriously condensed, but also partially reworked variant of that part of the database, with all the words rearranged in order of their relative historical stability.

First I discuss the subset of «super-stable items» that have been retained from EOC all the way to PTH (this is the largest sub-set, but also understandably requiring the least amount of commentary); then group B consists of «medium-stable items», for which it makes sense to postulate one replacement over the analyzed 2,500-year long period; finally, the shortest and the most difficult group C consists of «highly unstable» items that may have undergone no fewer than two replacements over the same period. Group D lists two interesting deviations where intermediate periods may show «dead-end» dialectal semantic developments, and, finally, Group E lists one item that has been excluded from analysis due to insufficient data.

A. Super-stable items (61 words).

A.1. Items attested with the same root morpheme throughout all four stages of Chinese.

A.1.1. 'big': 大 (**dhaːts* > *dà*).

A.1.2. 'black': 黑 (**s=məːk* > *hēi*). ◊ Transparently derived from 墨 **məːk* > *mò* 'ink', but still clearly the primary neutral equivalent for 'black' already in EOC. The idea that 黑 *hēi* had replaced the earlier 玄 **gʷiːn* > *xuán* in this meaning during the Zhōu period (Schuessler 2007: 277) seems to rely more on the derived origin of *hēi* than concrete textual evidence: there are, in fact, no contexts at all in EOC or COC literary monuments where *xuán* should be unambiguously translated as 'black' rather than a more general 'dark'[4]. For a good context supporting a basic function for 黑 *hēi* (as well as 赤 *chì* 'red', see below), cf. 莫赤匪狐莫黑匪烏 *mò chì fěi hú, mò hēi fěi wū* «there is nothing redder than a fox, nothing blacker than a raven» (*Shījīng* 41, 3); no such diagnostic contexts are available for 玄 *xuán* or any of the even more rare quasi-synonyms for 'black, dark', such as 緇 *zī* (only found twice in the *Shījīng* applied to some names for garments).

A.1.3. 'blood': 血 (**swiːt* > *xuè*).

A.1.4. 'cloud': 雲 (**whən* > *yún*).

A.1.5. 'come': 來 (**rəː* > *lái*).

A.1.6. 'die': 死 (**siyʔ* > *sǐ*).

A.1.7. 'dry': 乾 (**ghar* > *gān*).

A.1.8. 'ear': 耳 (**nhəʔ* > *ěr*). ◊ In the modern language, used primarily as part of the binome 耳朵 *ěr-duǒ*, lit. 'ear-cluster'.

A.1.9. 'fire': 火 (**səməːyʔ* > *huǒ*).

A.1.10. 'fish': 魚 (**ŋha* > *yú*).

A.1.11. 'hair /of head/': 髮 (**pat* > *fá*). ◊ All four stages of Chinese show a very clear and persistent lexical differentiation between **pat* 'hair of the head' (in the modern language, typically used as part of the binome 頭髮 *tóu-fá* 'head-hair') and 毛 **mhaːw* 'hair on the body' (also 'wool', 'fur', etc.).

A.1.12. 'hand': 手 (**tlhuʔ* > *shǒu*).

[4] A different opinion is voiced in Wu 2011: 87, where it is stated that in the corpus of bronze inscriptions, 玄 *xuán* is more frequent than 黑 *hēi* and is a better candidate for «basic 'black'» than the latter. However, Wu does not list any diagnostic contexts; frequency alone is not a clinching argument here, if, for instance, 玄 *xuán* (like 緇 *zī* in later received texts) was a typical term for denoting specific shades of ceremonial clothing, frequently depicted in bronze inscriptions. Note that most of our other observations on the evolution of color terms largely coincide with the thorough analysis presented in Wu 2011.

A.1.13. 'heart': 心 (**səm* > *xīn*). ◊ In the modern language, used primarily as part of the binome 心臟 *xīn-zàng*, literally 'heart-store'. Already in the ancient texts, the word is much more frequently found in abstract meanings ('mind', 'soul', 'conscience', 'intention', etc.) than in the required anatomical meaning; however, there is no evidence whatsoever that Chinese ever knew a different term for the anatomical 'heart'.

A.1.14. 'horn': 角 (**kroːk* > *jiǎo*).

A.1.15. 'I': 我 (**ŋhaːyʔ* > *wǒ*). ◊ For the EOC period, 予 ~ 余 (**dla* > *yú*) must be added as a synonym; the semantic difference between *wǒ* and *yú* is a much debated and still unresolved issue. However, both variants are known already from the Shang period, so there are no arguments in favor of a lexical replacement (merely the elimination of one of the synonyms in the COC period). In COC as well as in certain series of Zhōu epigraphic inscriptions, 我 **ŋhaːyʔ* co-exists with the morphological variant 吾 **ŋha*, but this has no bearing on lexicostatistical calculations, since the root morpheme is obviously the same.

A.1.16. 'kill': 殺 (**sraːt* > *shā*). ◊ There are some signs that in the modern language, the old word 殺 *shā* (or its bisyllabic counterpart 殺死 *shā-sǐ*) is being gradually replaced by the colloquial 打死 *dǎ-sǐ* (lit. 'hit-die'), but 殺 *shā* is still a frequent and stylistically neutral equivalent.

A.1.17. 'know': 知 (**tre* > *zhī*). ◊ Typically used as part of the binome 知道 *zhī-dào* in the modern language. It is useful to note that in the *Línjì lù* dialect this word is in free competition with the synonymous 識 (**tək* > *shì*), whose meaning in COC is closer to 'learn, keep in memory' and in the modern language to 'be acquainted with smbd.'; cf. contexts such as 總識伊來處 «[I] always know the place from which he comes», etc. However, this observation has no impact on the overall statistics for lexical replacements.

A.1.18. 'leaf': 葉 (**lhap* > *yè*). ◊ Extended with the desemanticized suffix 子 in the modern language (葉子 *yè-zi*).

A.1.19. 'many': 多 (**taːy* > *duō*).

A.1.20. 'meat': 肉 (**nhuk* > *ròu*).

A.1.21. 'moon': 月 (**ŋot* > *yuè*). ◊ Typically used as part of the binome 月亮 *yuè-liàng* (lit. 'moon-shine') in the modern language.

A.1.22. 'mountain': 山 (**sraːn* > *shān*).

A.1.23. 'name': 名 (**mheŋ* > *míng*). ◊ Typically used as part of the binome 名字 *míng-zì* (lit. 'name-cognomen') in the modern language.

A.1.24. 'new': 新 (**sin* > *xīn*).

A.1.25. 'night': 夜 (**lias* > *yè*).

A.1.26. 'nose': 鼻 (**bhits* > *bí*). ◊ Extended with the desemanticized suffix 子 in the modern language (鼻子 *bí-zi*).

A.1.27. 'not': 不 (**pə* > *bù*).

A.1.28. 'one': 一 (**ʔit* > *yī*).

A.1.29. 'person': 人 (**nin* > *rén*).

A.1.30. 'rain': 雨 (**whaʔ* > *yǔ*).

A.1.31. 'see': 見 (**keːns* > *jiàn*).

A.1.32. 'sit': 坐 (**ʒoːyʔ* > *zuò*). ◊ The word is only scarcely attested in EOC, and there may be some doubt as to whether it was really the most common and neutral equivalent for 'sit' during that period; a possible competitor is 居 (**ka* > *jū*, with a possible falling tone variant **ka-s*) 'to stay, dwell, reside', for which some contexts might suggest an earlier semantics of 'sit'. There are, however, no strong arguments for taking 坐 *zuò* out of the lexicostatistical comparison; at best, 坐 *zuò* and 居 *jū* could be thought of as synonyms (for the EOC stage only).

A.1.33. 'small': 小 (**sewʔ* > *xiǎo*). ◊ Several more specific adjectives denoting minuscule size are found in the texts (e.g. 細 **seːs* > *xì*, 微 **məy* > *wéi*), but they are statistically infrequent and

never feature in the standard antonymous pair 大 *dà* 'big' vs. 小 *xiǎo* 'small', for which there are multiple examples in the *Shījīng*.

A.1.34. 'stone': 石 (**diak* > *shí*). ◊ Usually extended with the desemanticized suffix 頭 in the modern language (石頭 *shí-tou*).

A.1.35. 'swim': 游 (**lu* > *yóu*). ◊ In the *Línjì lù*, only attested in application to fish (遊魚何得迷 'how did the fish that swim lose their way?'), but no evidence for any different verb denoting the corresponding human activity. In the modern language, mainly used as part of the binome 游泳 *yóu-yǒng*, where 泳 *yǒng* (attested already in the *Shījīng*) seems to be the original equivalent for 'to wade (in water)'.

A.1.36. 'tail': 尾 (**məyʔ* > *wěi*). ◊ Extended with the desemanticized component 巴 *bā* (etymologically = 把 *bǎ* 'handle') in the modern language (尾巴 *wěi-bā*).

A.1.37. 'thou': 汝 (**nhaʔ* > *rǔ*) ~ 爾 (**nheyʔ* > *ěr*). ◊ Both of these variants (freely interchangeable in some texts, dialectally or syntactically conditioned in others), as well as the modern variant 你 *nǐ*, clearly go back to the same root; alternations in the coda sometimes reflect archaic morphology and sometimes irregular dialectal developments, understandable for such high frequency usage forms as personal pronouns. No lexical replacements identified.

A.1.38. 'tongue': 舌 (**lat* > *shé*). ◊ Typically used as part of the binome 舌頭 *shé-tou* (with the same desemanticized suffix as in 'stone' q.v.) not only in the modern language, but already in MC: both the short variant *shé* and the disyllabic form are encountered in the *Línjì lù* as free variants.

A.1.39. 'warm (hot)': 熱 (**ŋet* > *rè*). ◊ For this entry, we choose 'hot' (= 'exceeding tolerable temperature') rather than 'warm', as allowable in the GLD. Unlike 'warm' (OC 溫 **ʔūn* > *wén*; modern 暖 *nuǎn*), 'hot' is quite stable throughout all four stages of Chinese.

A.1.40. 'water': 水 (**tuyʔ* > *shuǐ*).

A.1.41. 'we': 我 (**ŋhaːyʔ* > *wǒ*). ◊ In EOC and COC, sg. 'I' and pl. 'we' were usually not distinguished from each other. From Hàn times on, the differentiation, when necessary, is performed by desemanticized quasi-suffixal morphemes (我公 *wǒ-gōng*, 我等 *wǒ-děng*, 我們 *wǒ-men* etc.) without any replacements for the root morpheme.

A.1.42. 'white': 白 (**braːk* > *bái*).

A.1.43. 'who': 誰 (**duy* > *shuí*). ◊ The morphological derivate 孰 **du-k* (> *shú*), originally 'which one /out of several/?', sometimes replaces the original 誰 *shuí* in some dialects of late OC, but this has no bearing on the overall statistics.

A.1.44. 'woman': 女 (**nraʔ* > *nǚ*). ◊ Used by itself or within the binome 女人 *nǚ-rén* (lit. 'woman-person') in the modern language.

A.1.45. 'yellow': 黃 (**ghʷaːŋ* > *huáng*).

A.2. Items not attested in the *Línjì lù* dialect of MC, but well attested at the three other stages.

A.2.1. 'bird': 鳥 (**tiːwʔ* > *niǎo*). ◊ Initial *n-* in the Běijīng dialect is irregular, but the word is still clearly cognate with its OC predecessors. Should be distinguished from OC 禽 **ghəm* 'game-bird', used mainly in hunting contexts.

A.2.2. 'fat': 脂 (**kiy* > *zhī*). ◊ In the modern language, mainly used as part of the binome 脂肪 *zhī-fáng* (already attested in texts going back to the Jìn dynasty, 3rd–5th centuries A.D.). For both stages of OC, an additional synonym is the word 膏 **kāw* (> *gāo*); semantic difference between **kiy* and **kāw* is impossible to reliably determine based on the available text corpus (in the *Shuōwén jiězì* **kiy* is explained as 'fat of horned cattle' and **kāw* as 'fat of hornless cattle' — an explanation not explicitly confirmed by textual usage, but showing that the two words must have been very close). However, 脂 **kiy* is well attested already in the *Shījīng*, and the existence of an additional synonym is not a reason for postulating a lexical replacement.

A.2.3. 'feather': 羽 (**w/r/aʔ > yǔ*). ◊ In the modern language, normally used as part of the binome 羽毛 *yǔ-máo*, lit. 'feather-hair'.

A.2.4. 'fly /v./': 飛 (**pəy > fēi*).

A.2.5. 'long': 長 (**draŋ > cháng*).

A.2.6. 'round': 圓 (written simply as 員 in the earlier texts; **wran > yuán*). ◊ Attestation in the adjectival meaning in EOC and early COC is extremely scarce and dubious, but verbal ('to be around') and nominal ('circumference') meanings are attested (Schuessler 1987: 791), and there are no other serious candidates for the expression of the adjectival meaning in those periods.

A.2.7. 'sand': 沙 (**sraːy > shā*).

A.2.8. 'seed': 種 (**toŋʔ > zhǒng*). ◊ Extended with the desemanticized suffix 子 in the modern language (種子 *zhǒng-zi*).

A.2.9. 'skin': 膚 (**pra > fū*). ◊ In the modern language, used only as part of the binome 皮膚 *pí-fū*, where 皮 (**bhay > pí*) is also a very old word, encountered much more frequently than **pra* already in EOC (Schuessler 1987: 169, 457); however, its EOC attestations are completely restricted to the notion of 'animal skin', 'fur', 'hide', transparently separating it from the required Swadesh meaning of 'human skin'. The first references to **bhay* as 'human skin' seem to appear no earlier than in Hàn-era texts, and even then mostly as part of the already attested binome 皮膚 *pí-fū* (co-existing with simple *fū*).

A.2.10. 'star': 星 (**sheːŋ > xīng*).

A.3. Items not attested (properly) in EOC, but stable throughout all other periods.

A.3.1. 'ashes': COC 灰 (**smaːy > huī*). ◊ Not attested at all in EOC (nor in the *Línjì lù*, for that matter), but this is the only word with the basic meaning 'ashes' throughout the entire known history of Chinese. Even the graphic shape of the character ('hand' + 'fire') suggests an archaic origin, despite not being attested in epigraphic monuments.

A.3.2. '/tree/-bark': COC 皮 (**bhay > pí*). ◊ It seems that the basic root for 'tree-bark' has always been the same as the root for '/animal/ skin, hide' in general (see A.2.9), although specific instances of 'bark' are lacking in both EOC and the *Línjì lù*. In the modern language, the default equivalent is rather the binomial 樹皮 *shù-pí*, where 樹 *shù* = 'tree'; this does not count as a replacement.

A.3.3. 'bone': COC 骨 (**kuːt > gǔ*). ◊ Strangely enough, the word 'bone' is not at all attested in EOC; however, the graphic shape of the character looks archaic, and there is no specific reason to suggest that the EOC equivalent may have been different. In the modern language the word is usually extended with the desemanticized suffix 頭 (骨頭 *gǔ-tou*).

A.3.4. 'knee': COC 膝 (**sit > xī*). ◊ A somewhat problematic entry; the word 'knee' is not really attested in Chinese until texts typically dated to around the 3rd — 1st cent. BC (*Xún-zǐ*, etc.), nor is it encountered in the *Línjì lù*. Again, however, nothing indicates the existence of any other word in this meaning throughout all the stages of non-dialectal Chinese. In the modern language, the default equivalent is the binome 膝蓋 *xī-gài*, lit. 'knee-cover', that does not count as a replacement.

A.3.5. 'liver': COC 肝 (**kaːn > gān*). ◊ Well attested in COC (though not in early Confucian texts) and MC, but not found in EOC. No indication of any possible alternate equivalents throughout any of the stages of written Chinese.

A.3.6. 'louse': COC 蝨 (**srit > shī*). ◊ Attested in COC (though not in early Confucian texts), but not known in EOC or in the *Línjì lù*. Extended with the desemanticized suffix 子 in the modern language (蝨子 *shī-zi*). The word has a solid Sino-Tibetan etymology (= Tibetan *śig*, Lushai *hrik* 'louse' etc.), indirectly confirming that the word has been super-stable from the beginning.

B. Medium-stable items (31 words)

B.1. Replacements from EOC to COC.

B.1.1. 'breast (= chest)': EOC 膺 (**ʔrəŋ* > *yīng*) → COC 胸 (**sŋoŋ* > *xiōng*). ◊ The latter word is quite clearly the main equivalent for 'male chest' in both COC and the modern language, and is encountered once in the *Línjì lù* in the bound expression 指胸 *zhǐ-xiōng* 'to point at one's breast', which makes it at least a plausible candidate for the same meaning in MC. Conversely, the word is not encountered in any EOC texts, where the only known possible equivalent is 膺 **ʔrəŋ* (although it is largely used in bound expressions and figurative meanings as well). This is sufficient evidence to at least suspect a lexical replacement.

B.1.2. 'man': EOC 夫 (**pa* > *fū*) → COC 男 (**nəːm* > *nán*). ◊ A debatable choice. The assumed replacement **nəːm* is actually well attested already in EOC (Schuessler 1987: 436). However, throughout that period it is encountered infrequently, most often to denote a specific feudal title ('*nán*' = 'baron'); more basic usage is generally confined to the noun phrase 男子 **nəːm-cəʔ* '(male) son', used to specify the gender of the descendant (and thus opposed to 女子 **nraʔ-cəʔ* '(female) daughter'. Schuessler adds several epigraphic examples in which *nəːm* means 'male descendant, son' all by itself and may thus be an abbreviation of **nəːm-cəʔ* (e. g. 我後男 **ŋ^{h}aːyʔ g^{h}oːʔ nəːm* 'my (future) male descendants' [1381 Xuan], etc.). On the other hand, EOC **pa* is statistically far more frequent, and in most contexts, applied to human beings that are male by default (soldiers, farmers, etc.) or expressly meaning 'husband'. It is interesting that in the sole known early literary context in which we encounter the noun phrase 夫人 **pa-nin* [*Shàngshū* 42, 9], it clearly refers to 'man' or 'men', whereas already in COC the term **pa-nin* is more commonly used to denote the wife, i. e. 'man's person', rather than 'man-person'. As for the use of **pa* itself in the COC period, most texts clearly show that it is employed in a «socially marked» manner, either in the derived meaning 'teacher, master' (usually within the compound 夫子 **pa-cəʔ*), or in the meaning 'husband' (often within the antonymous pair 夫婦 **pa-bəʔ* 'husband(s) and wife (wives)'). All of this speaks in favor of a gradual transition from **pa* to **nəːm*, with **pa* still functioning as the main word for 'male person' in Early Zhōu.

B.1.3. 'road': EOC 道 (**lhuːʔ* > *dào*) → COC 路 (**raːks* > *lù*). ◊ In EOC, **lhuːʔ* is the most statistically frequent word denoting the idea of 'road' without any further connotations. It also serves as the basis for the derived verb 導 **l^{h}uː-s* 'to lead, conduct (along the way)' (Schuessler 1987: 116). The word 路 **raːks* 'road' (Schuessler 1987: 395), in comparison, is encountered only in a tiny handful of contexts, most often, within the noun phrase 路車 **raːks kla* 'grand chariot', where it is not even certain that the *raːks* in question represents the same 'word'. It is likely that the gradual replacement of **l^{h}uːʔ* with *raːks* did not really start until COC, possibly caused by the expanding polysemy of the former ('road / way / manner / habit / Tao', etc.).

In COC, the simple word 道 **l^{h}uːʔ* is very rarely employed to denote a physical 'road' by itself — most of the time, it only appears within the compound form 道路 **l^{h}uːʔ-raːks*. On the other hand, 路 **raːks* is very common as 'road' on its own, quite unlike its functions in the EOC period. Likewise, in the modern language the basic equivalent for 'road' is either the bisyllabic 道路 *dào-lù* or the monosyllabic 路 *lù*, but never the monosyllabic 道 *dào*. This fairly transparent shift in usage may count as a lexical replacement, with the original **l^{h}uːʔ* ceding its basic functions to **raːks*.

B.1.4. 'root': EOC 本 (**pəːrʔ* > *běn*) → COC 根 (**kəːn* > *gēn*). ◊ Although the absolute majority of contexts in which **pəːrʔ* is encountered in EOC are metaphorical ('root' as 'foundation', etc.), at least one context [*Shījīng* 255, 8] clearly refers to *pəːrʔ* as 'tree root', opposed to 枝 **ke* 'branches' and **lhap* 葉 'leaves'. The simple pictographic nature of the character also hints at the original semantics of 'tree root'. No other words with this meaning are found in EOC. By contrast, it cannot be doubted that by the end of COC the word 根 **kəːn* had completely replaced

the earlier **pəːrʔ* in the basic meaning 'root (of trees and other plants)', with **pəːrʔ* preserved in a wide range of figurative meanings ('root' as 'origin', 'foundation', 'essentials', etc.). In the *Shuōwén jiězì*, for instance, all of the references to roots of plants always comprise **kəːn*, whereas **pəːrʔ* is reserved for the more abstract meaning 'foundation'.

The difficult problem is to determine the approximate period during which the replacement actually took place. Early Confucian texts offer little help in this matter, since the word 'root' is only encountered in them in figurative meanings ('origin', most of the time), thus, only **pəːrʔ* is attested, but none of the attestations are diagnostic. Cf., however, a diagnostic context in the *Inner Chapters* of *Zhuāngzǐ* [1, 4, 6], a document of comparable antiquity: 仰而視其細枝... 俯而見其大根 *yǎng ér shì qí xì zhī... fǔ ér jiàn qí dà gēn* «he looked up and saw its (the tree's) thin branches... he looked down and saw its big roots». In light of all available evidence, we fill the COC slot with **kəːn*. In the modern language, the situation persists (although the root 根 *gēn* is typically used in binomial constructions, such as *shù-gēn* 樹根 'tree-root', etc.).

B.2. Replacements from COC to MC.

B.2.1. 'belly': OC 腹 (**puk* > *fù*) → MC 肚 ([*dó*] > *dǔ*). ◊ The new word for 'belly' is attested already in the *Línjì lù*: *lǜ niú dǔ lǐ shēng* 驢牛肚里生 '/you/ will be born in the belly of a donkey or a cow'. The new word persists in the modern language, albeit usually extended with the desemanticized suffix 子 (肚子 *dǔ-zi*).

B.2.2. 'burn (tr.)': OC 焚 (**bən* > *fén*) → MC 燒 (**sŋew* > *shāo*). ◊ In EOC, **bən* is the main word for 'burn' and **sŋew* is not attested at all. The latter appears in COC and gradually replaces the former as the most neutral equivalent for the concept: of note may be the statistical observation that in the *Zuǒzhuàn* (5th century BC) we observe 42 cases of **bən* vs. no cases at all of **sŋew*, but in the *Shǐjì* (1st century BC) we already see just 17 cases of **bən* vs. 58 cases of **sŋew* (sporadically, the compound form 焚燒 **bən-sŋew* is also observed). In the *Línjì lù*, the equivalent is either the compound form (e.g. *fén-shāo jīng xiàng* 焚燒經像 «to burn writings and images») or the simple 燒 *shāo* (*bèi huǒ lái shāo* 被火來燒 «you will be burned by fire»); the same situation is typical of the modern language. We may tentatively conclude that **bən* was essentially replaced by **sŋew* around Hàn-era times, i.e. in the interim period between COC and MC.

B.2.3. 'cold': OC 寒 (**gaːn* > *hán*) → MC 冷 (**reːŋʔ* > *lěng*). ◊ The word **reːŋʔ* 'cold' frequently appears in Hàn-era texts, but not in COC, where **gaːn* is still the default equivalent. By MC times, **gaːn* is clearly a bound and archaic form (in the *Línjì lù*, it is only encountered in the idiomatic collocation 寒松 *hán sōng* 'winter pine'), and it remains a bound form in the modern language.

B.2.4. 'eat': OC 食 (**ləkˀ* > *shí*) → MC 喫 (**kheːkˀ* > *chī*). ◊ An early colloquialism attested already in the *Shuōwén jiězì*, *chī* is transparently the neutral equivalent of the meaning 'eat' in the *Línjì lù* (*shí* and *chī* are both attested in the text, but only the latter is regularly encountered in direct speech, e.g. *yī rì chī duō shǎo* 一日喫多少 «how much do they eat per one day?»).

B.2.5. 'eye': OC 目 (**muk* > *mù*) → MC 眼 (**ŋrəːnʔ* > *yǎn*). ◊ The original meaning of the word may have been 'eye-ball' (although already in the *Shuōwén *ŋrəːnʔ* is explained as 目 **muk* 'eye'). In any case, the replacement is quite transparent in the *Línjì lù*, where the old word 目 **muk* is only encountered in bound expressions such as 目前 *mù-qían* 'present', etc.

B.2.6. 'head': OC 首 (**sluʔ* > *shǒu*) → MC 頭 (**dhoː* > *tóu*). ◊ This replacement may have already taken place in Hàn-era time (in the *Shǐ jì*, the word seems to be more frequent than *shǒu*, particularly in direct speech).

B.2.7. 'smoke': OC 熏 (**hun* > *xūn*) → MC 煙 (**ʔiːn* > *yán*). ◊ Available attestations are insufficient to reconstruct a completely reliable picture. The facts so far are as follows: (a) only **hun* is attested in EOC; (b) **ʔiːn* is clearly the main equivalent for 'smoke' in all Hàn-era and later texts; (c) early Confucian texts of the 5th-6th centuries have no occurrences of 'smoke', but the

word is sometimes encountered in texts such as *Mò-zǐ* or *Zhuāng-zǐ*, albeit more often in the verbal ('to smoke out') than nominal meaning. We tentatively assume that the replacement of the original noun has to be dated to a time period around Early Hàn, but new data may overturn this assumption.

B.2.8. 'tree': OC 木 (**mhoːk* > *mù*) → MC 樹 (**dhoʔ* > *shù*). ◊ The nature and reasons for this replacement are quite transparent: it begins as a compound form 樹木 *shù-mù*, lit. 'planted tree' (where 樹 = 豎 **dhoʔ/s/* 'plant vertically'), well attested already in the Hàn period. By late MC, the replacement seems to be complete: in the *Línjì lù*, simple 樹 *shù* is the usual equivalent for 'tree /growing/' (cf. *chéng yī zhū dà shù* 成一株大樹 «he will become a big tree»), while 木 *mù* is restricted to the meaning 'wood /material/'. In the modern language, 'tree' is 樹 *shù* or 樹木 *shù-mù*; 木 *mù* (more frequently, the extended suffixal variant 木頭 *mù-tóu*) is strictly 'wood'.

B.2.9. 'two': OC 二 (**niːs* > *èr*) → MC 兩 (**rhaŋʔ* > *liǎng*). ◊ This only counts as a replacement if we follow the definition of 'two' as an adjectival lexeme, used in conjunction with a quantified noun; since this is the most common function of numerals, such a definition is, however, fully justified. The replacement process is well traceable across ancient texts. The word **rhaŋʔ* is not encountered at all as a numeral in EOC texts; is rigidly restricted to paired objects only throughout COC (兩手 *liǎng shǒu* 'two hands', 兩馬 *liǎng mǎ* 'a pair of horses' etc.); and begins to be freely applied to any objects, paired or not, around Hàn times. In the *Línjì lù* it is clearly the same default equivalent for 'two /of anything/' as it is in the modern language, e.g. 與爾兩文錢 *yǔ ěr liǎng wén qián* 'I give you two coins', etc.

B.2.10. 'go (walk)'[5]: OC 往 (**waŋʔ* > *wǎng*) → MC 去 (**khas* > *qù*). ◊ This replacement is rather tricky and not easily detectable through the corpus, particularly considering the general abundance of verbs denoting directed movement in OC (partial synonyms also include 之 **tə* 'to go, be headed somewhere', 適 **tek* 'to go', etc.). Nevertheless, it can be more or less ascertained that throughout EOC and COC 去 *qù* is almost exclusively used in the meaning 'to /take/ leave', and, even more importantly, that the basic antonymous pair 'come and go' is always rendered as 往來 *wǎng-lái* rather than 往去 *wǎng-qù*. This situation is completely reversed in the language of the *Línjì lù*, where the usual antonym of 來 *lái* is always 去 *qù* rather than 往 *wǎng*, and remains as such in the modern language.

B.2.11. 'what': OC 何 (**ghaːy* > *hé*) → MC 什麼 ([*ʒimmu̯a*] > *shémme*). ◊ While the old inanimate interrogative pronoun still survives in MC as an archaism or as part of some bound expressions, it is clear that already in the *Línjì lù* the default equivalent is the replacement shémme, a colloquialism that arose already in post-Hàn times.

B.3. Replacements from MC to PTH.

B.3.1. 'nail (claw)'[6]: OC 爪 (**cruːʔ* > *zhǎo*) → PTH 指甲 *zhǐ-jiǎ*. ◊ In the *Línjì lù*, the old word 爪 *zhǎo* still seems to be the default equivalent, cf. 髮毛爪齒 *fá-máo zhǎo chǐ* «head hair, body hair, nails, and teeth». The binome 指甲 *zhǐ-jiǎ* (literally 'finger-shell') is first attested in Sòng-era texts (11–12 cent.).

B.3.2. 'give': OC 畀 (**pits* > *bì*) / 予 ~ 與 (**laʔ* > *yǔ*) → PTH 給 *gěi*. ◊ In EOC, **pits* and **laʔ* are basically interchangeable synonyms, cf. two lines in the same *Shījīng* poem (53, 1): 何以畀之 *hé yǐ bì zhī* vs. 何以予之 *hé yǐ yǔ zhī*, both translatable as 'what shall I give him?' Only the latter, however, survives into COC times, where it becomes the sole neutral equivalent for the re-

[5] The meaning 'go' (i.e. the opposite of 'come') is consistently used in the Global Lexicostatistical Database instead of 'walk' (i.e. 'move without a specific direction') in the «classic» Swadesh wordlist, but is still filed alphabetically under 'walk' because of technicalities.

[6] The meaning '(finger)nail' (of human) is consistently used in the Global Lexicostatistical Database instead of 'claw' (animal) in the «classic» Swadesh wordlist, but is still filed alphabetically under 'claw' because of technicalities.

quired meaning and persists into MC. PTH 給 *gěi* is a more recent replacement (a dialectal phonetic development from MC *kip* ← OC **kəp*, originally 'to provide, furnish').

B.3.3. 'green': OC 青 (**sheːŋ* > *qīng*) → PTH 綠 *lǜ*. ◊ Both these words are already attested in EOC and persist all the way to the modern language. Our decision is based primarily on diagnostic contexts, such as the application of these qualifiers to specifically green objects (e.g. 'leaves') and their appearance in lists of the most basic color terms. The latter, in particular, allows to assume that 青 *qīng* was still the basic 'green' as late as MC (cf. in the *Línjì lù*: 把我著底衣, 認青黃赤白 *bǎ wǒ zhuó-di yī, rèn qīng huáng chí bái* «he seizes the clothes that I wear, considers them to be green, yellow, red or white»). In the modern language, however, 青 *qīng* has shifted to denote a darker tinge of green, with 綠 *lǜ* taking its place in the general spectrum.

B.3.4. 'hear': OC 聞 (**mən* > *wén*) → PTH 聽見 *tīng-jiàn*. ◊ The old word is still the default equivalent for 'hear' in the *Línjì lù*; in the modern language, it is only encountered in bound expressions.

B.3.5. 'mouth': OC 口 (**khoːʔ* > *kǒu*) → PTH 嘴 *zuǐ*. ◊ The latter word, originally written simply as 觜, used to mean 'beak'; the shift to 'mouth' is apparently a very recent development that took place sometime in the late Qíng period.

B.3.6. 'red': OC 赤 (**khiak* > *chì*) → PTH 紅 *hóng*. ◊ The latter word is already attested in COC, where it, however, is very rare and most likely denotes some specific shade of red. 赤 *chì* is still the main equivalent for 'red' in the *Línjì lù* (see the example in B.3.2). It is not quite clear at which particular moment the replacement has become complete, but in the modern language 赤 *chì* is no longer in active usage. Other OC words that are typically translated as 'red', e.g. 朱 *zhū*, 彤 *tóng*, etc., are statistically less frequent and more commonly found in conjunction with articles of clothing than natural objects.

B.3.7. 'stand': OC 立 (**rəp* > *lì*) → PTH 站 *zhàn*. ◊ The older meaning of 站 *zhàn* is 'to stop somewhere; to occupy a place' (originally written as 佔). The word gradually replaces the older 立 *lì* in the basic meaning 'to stand' over the Míng-Qíng period.

B.3.8. 'sun': OC 日 (**nit* > *rì*) → PTH 太陽 *tài-yàng*. ◊ The metaphoric term 太陽 *tài-yàng*, lit. 'the extreme Yang', is well attested since at least Hàn times, but only functions as the default term for the celestial body in the modern language.

B.3.9. 'this': OC 此 (**cheyʔ* > *cǐ*) → PTH 這 *zhè*. ◊ There is a certain number of stems that may be used to denote proximal deixis at any given time period in Chinese, but 此 *cǐ* is the one link that ties together all these time periods — with the exception of the modern language, where it is only used in idiomatic bound forms, while the common equivalent for 'this' is the replacement 這 *zhè*. In the *Línjì lù*, both 此 *cǐ* and 這 *zhè* co-exist, but 此 *cǐ* is still far more common and cannot be formally regarded as a literary archaism.

B.3.10. 'tooth': OC 齒 (**thəʔ* > *chǐ*) → PTH 牙 *yá*. ◊ The story here is as follows: (a) in EOC, 齒 *chǐ* = 'teeth /of humans or animals/', 牙 *yá* = '/special/ teeth /of animals only/' (usually tusks, possibly also fangs etc., i.e. protruding teeth; even the graphic shape of the character suggests 'tusks'); (b) in COC, the situation is largely the same, although in a few cases the compound form 齒牙 *chǐ-yá* is also attested; (c) in the *Línjì lù*, the usual equivalent is either bisyllabic 牙齒 *yá-chǐ* or monosyllabic 齒 *chǐ*, but never monosyllabic 牙 *yá*; (d) conversely, in the modern language, the usual equivalent is either bisyllabic 牙齒 *yá-chǐ* or monosyllabic 牙 *yá*, but never monosyllabic 齒 *chǐ*. According to our rules, this indicates a replacement from MC to PTH.

B.4. Unclear due to lack of attestation in MC.

B.4.1. 'dog': OC 犬 (**khʷiːnʔ* > *quǎn*) → PTH 狗 *gǒu*. ◊ Although the word 'dog' is not attested in the *Línjì lù*, it may be reasonably well guessed that 狗 *gǒu* had already become the primary equivalent for the neutral meaning 'dog' in MC, judging by the steady increase in at-

testation since Hàn times, by which period the old 犬 *quǎn* had largely been demoted to the specialized meaning 'hunting dog = hound'. See Starostin 2013 on the possible semantic differentiation between *quǎn* and *gǒu* in COC (where *gǒu* may have originally denoted a special breed of dogs raised for meat).

B.4.2. 'drink': OC 飲 (**ʔəmʔ* > *yǐn*) → PTH 喝 *hē*. ◊ Not attested in the *Línjì lù* at all. The modern equivalent 喝 *hē* is only encountered in texts since the Yuán dynasty (13th — 14th centuries), so it may be assumed that the old word was still in colloquial circulation throughout the MC period.

B.4.3. 'egg': COC 卵 (**rhoːnʔ* > *luǎn*) → PTH 蛋 *dàn*. ◊ The old word is not attested either in EOC (although the pictographic nature of the character may suggest an archaic origin) or in the *Línjì lù*. The new word is a transparent semantic extension of *dàn* 'ball, pill, bullet, any small round object', a word well attested already in OC and usually written as 彈 in its original meaning. The first attestations of the semantic shift come from classic 16th–18th century novels; it may be assumed that the old word *luǎn* was still the basic term in MC[7].

B.4.4. 'full': OC 盈 (**leŋ* > *yíng*) → PTH 滿 *mǎn*. ◊ Not attested in the *Línjì lù*. The original meaning of 滿 *mǎn* was likely 'to fill up, overflow (of water)'; it is not found in the generic meaning 'to fill /anything/' or in the adjectival meaning 'full' in early Confucian texts or in the *Dàodéjīng*, but is already competing with 盈 *yíng* in *Zhuāngzǐ*. In the *Shījì*, 盈 *yíng* is encountered 14 times next to 85 for 滿 *mǎn*, meaning that the replacement was likely complete by the early Hàn period.

Another semantically close morpheme, 充 (**thuŋ* > *chōng*), is first encountered in the *Shījīng* as part of the compound noun 充耳 *chōng-ěr* 'ear stopper'; in COC it is usually applied to the process of filling up storage units (granaries, etc.) and also used in various figurative meanings. The bisyllabic compound 充滿 *chōng-mǎn* is well attested already in Early Hàn times and has persisted all the way up to modern times; nevertheless, 充 *chōng* almost always behaves as a secondary morpheme in this formation, and while it is hard to precisely state the semantic difference between *chōng* and *mǎn* in the COC period (it may have been 'to fill up with hard substances' vs. 'to fill up with liquid substances', as one of the possibilities), including it in our calculations as a secondary synonym or excluding it altogether will have no effect on the overall calculations.

B.4.5. 'neck': OC 領 (**rheŋʔ* > *lǐng*) → PTH 脖子 *bó-zi*. ◊ Not attested in the *Línjì lù*. Modern *bó-zi* is a very late word, not attested earlier than the Yuán dynasty (13th–14th centuries). In addition, a very frequent equivalent for 'neck' in early Hàn texts is OC 項 **groːŋʔ* (> *xiàng*), whereas 領 is more frequently used in the meaning 'collar' by that time. It cannot, however, be confirmed at this time that 項 *xiàng* continued to be the main term for 'neck' throughout MC. Another occasional synonym in COC is 脰 (**dhoːs* > *dòu)*, always translated as 'neck'; in about 90% of its occurrences in texts, it is used as the object of 'breaking' or 'cutting', implying immediate death, so it is possible that a more exact meaning is something like 'neck vertebra'. In any

[7] It is suggested in Baxter, Sagart 2014: 324 that a more archaic equivalent for 'egg' may be a root **tʰu[n]* (= **tʰun* or **tʰur*), not attested in any written Chinese texts but functioning as a vulgar equivalent for 'egg' and/or 'testicles' in some Southern dialects (Cantonese *tʃʰœn*¹, Hakka *tʃʰun*¹); its antiquity is allegedly corroborated by semantically and phonetically perfect Tibeto-Burman parallels. Regardless of whether this hypothesis is correct, it could only be taken into consideration in this paper if we were to assert that this **tʰu[n]*, not 卵, had the basic meaning 'egg' in EOC, and that somehow Cantonese and Hakka had managed to inherit it, completely bypassing the COC and MC stages. Since the first part of this statement has no confirmation in written evidence and the second is almost impossible to believe, at best we could hypothesize that **tʰu[n]* may have existed in EOC and COC side-by-side with 卵 as a «vulgar» synonym, managing to survive into Cantonese and Hakka; but this hypothesis would have no bearing on our lexicostatistics, which requires that only the stylistically neutral equivalents be taken into consideration.

case, it is a statistically infrequent (no more than a couple dozen entries in the entire COC + Hàn corpus, next to hundreds for 領 **rheŋʔ* and 項 *xiàng*) and contextually bound word.

B.4.6. 'that': OC 彼 (**payʔ* > *bǐ*) → PTH 那 *nà* ~ *nèi*. ◊ Not attested in the *Línjì lù*, although apparently certain other texts in the *yǔlù* genre already show 那 *nà* as the basic adjectival stem denoting objects that are far away, while 彼 *bǐ* is more frequently restricted to adverbial functions ('there', 'in that place'). On the other hand, cf. B.3.9 'this' where it can be seen that both the old and the new pronoun still co-exist in the *Línjì lù* dialect as synonyms; it cannot be excluded that the same situation was symmetrically relevant for the distal deixis pronouns.

C. Unstable items (5 words)

C.1. EOC → COC, COC → MC.

C.1.1. 'bite': EOC 啑 (**di:t* > *dié*) → COC 噬 (**dats* > *shì*) → MC 咬 (**ŋhra:wʔ* > *yǎo*). ◊ The double replacement is quite uncertain[8]: so far, the only unambiguous EOC context with the verb 'to bite' is a passage in the earliest layer of the *Yìjīng*: 履虎尾, 不啑人 *lǚ hǔ wěi, bù dié rén* «if one steps on a tiger's tail, he does not bite». The situation in COC is also far from clear: statistically and contextually, there is some serious competition for 噬 **dats* on the part of 齧 (**ŋhe:t* > *nié*), also encountered several times (*Zhuāng-zǐ*; *Guǎn-zǐ*) in the meaning 'to bite' (or perhaps 'to gnaw?') as applied to dogs. The distinction between **dats* and *ŋhe:t* may have originally been dialectal (e. g. «Northern» vs. «Southern»), but it becomes seriously blurred in Hàn times (thus, both terms are interchangeable in the *Huáinán-zǐ*). Since MC, however, 咬 *yǎo* seems to have largely stabilized as the primary equivalent for this meaning.

C.1.2. (?) 'foot': EOC 趾 (**təʔ* > *zhǐ*) → COC 足 (**cok* > *zú*) → MC 腳 (**kak* > *jiǎo*). ◊ The fact that the 'foot' / 'leg' opposition in the earliest stages of Chinese was lexicalized as 趾 (originally written simply as 之) *zhǐ* 'foot' vs. 足 *zú* 'leg' is suggested, first and foremost, by the early graphical shapes of the characters:  'foot' vs.  'leg'. Textual evidence is ambiguous at best, since both 'feet' and 'legs' are very rarely attested in EOC, but at least one context in the *Shījīng* (麟之趾 *lín zhī zhǐ* 'the feet (= hooves) of the *lín*') indirectly supports this difference. In COC the old word *zhǐ* seems to have shifted its meaning to 'toe', while both 'foot' and 'leg' seem to merge into 足 *zú* for a while — at least until Hàn-era texts, when the differentiation re-emerges with the appearance of a new word for 'foot', 腳 *jiǎo* (not attested in EOC at all).

C.1.3. 'sleep': EOC 寐 (**miys* > *mèi*) → COC 臥 (**ŋho:ys* > *wò*) or COC 寢 (**shimʔ* > *qǐn*) → MC 睡 (**doys* > *shuì*). ◊ In EOC, 寐 **miys* is the most common designation of the static meaning 'sleep'; 寢 **shimʔ* is more rare and better interpreted as the dynamic 'lie down to sleep', or causative 'put to sleep' (antonymous to 興 *xīng* 'rise'). In COC, 寐 **miys* is practically non-existent, whereas 寢 **shimʔ* is sometimes found in unambiguously static contexts (e.g. 宰予晝寢 *zǎi yú zhòu qǐn* «Zai Yu slept during the day» [Lùnyǔ 5, 10]); however, it seems to be competing for the 'sleep' slot with 臥 **ŋho:ys*, a word that can be interpreted as 'to lie' or 'to sleep' depending on the context. By Hàn times, the word 睡 **doys* makes its appearance, and seems to completely eliminate all competition by the beginning of the MC period.

C.2. EOC → COC, MC → PTH.

C.2.1. 'all': EOC 率 (**srut* > *shuài*) or 咸 (**gro:m* > *xián*) → COC 皆 (**kro:y* > *jiē*) → PTH 都 *dōu*. ◊ We equate 'all' with the most commonly used Chinese adverbial adjuncts with the same meaning, typically placed right before the verb. EOC uses a variety of those, making it impos-

[8] Laurent Sagart (p.c.) has suggested the possibility of both **di:t* and **dats* reflecting the same original root, but the vocalism seems to go against this idea; even if this were so, the morphological alternation must have been so ancient that the two forms would hardly feel related in the 1st millennium BC.

sible to choose between **srut* and **grəːm*. In COC, 皆 **krəːy* is unquestionably the most widely used adjunct, although by early Hàn times it begins to compete with the synonymous 悉 (**sit* > *xī*); in the *Línjì lù*, **krəːy* is still encountered either on its own or in conjunction with **sit* (both 悉皆 *xī-jiē* and 皆悉 *jiē-xī* are possible). Curiously, modern 都 *dōu* seems to have already existed in its current meaning at least in Hàn times, but is only very occasionally attested until the modern phase of the language.

C.3. COC → MC, MC → PTH.

C.3.1. 'say': EOC 曰 (**wat* > *yuē*) → MC 云 (**wən* > *yún*) → PTH 說 *shuō*. ◊ We understand 'say' here as the most common verb to introduce direct speech, which makes it easier to single out one particular candidate among a huge variety of verbs denoting various kinds of speech in Chinese. In Old Chinese, this verb has always been 曰 **wat*; in the *Línjì lù*, direct speech is usually introduced by 云 **wən*, a verb already well attested in OC as well but nowhere near as common as **wat* (its functions in various subperiods and dialects are still somewhat unclear). In colloquial PTH, the functions of these words have been completely overtaken by 說 *shuō*, a word originally meaning 'to explain, interpret'.

D. Unusual deviations

These two cases describe interesting situations where one of the two intermediate attested stages features a variant that is deviant of the common form, so that older and newer forms of the language share the same equivalent but the intermediate equivalent is expressed by a different root.

D.1. 'earth': EOC 土 (**thaːʔ*) → PTH *tǔ* vs. MC 地 (*dî*). ◊ The semantic difference between 土 *tǔ* and 地 *dì* 'earth, ground' is often neutralized in both ancient and modern contexts, most obviously so within the compound formation 土地 *tǔ-dì*, well attested already in OC. Nevertheless, whenever the two morphemes are met separately, the former typically refers to 'earth' as substance ('soil' — the required Swadesh meaning) and the latter as surface ('ground', 'territory'). Surprisingly, one glaring exception is the dialect of *Línjì lù*, where it is 地 *dì* rather than 土 *tǔ* that commonly functions as a substance term, cf.: 被地水火風 *bèi dì shuǐ huǒ fēng* «suffer earth, water, fire, and wind» (the elements), etc., whereas the word 土 *tǔ* is almost always encountered only within the compound form 國土 *guó-tǔ* «territory (of state)». It is possible that this usage reflects a genuine case of lexical replacement in the respective dialect, though a specific peculiarity of the literary language is not excluded either.

D.2. 'good': EOC 好 (**huːʔ*) → PTH *hǎo* vs. COC 善 *danʔ* (→ PTH *shàn*). ◊ Curiously, the character 好 throughout most of the Classical Chinese period is most often employed to transcribe the derived verbal stem *huː-h* 'to love' rather than the original adjectival stem *huːʔ* 'good' (as in EOC); the latter cannot by any means pretend to denote the basic qualitative predicate '(to be) good' in any of the early Confucian texts or, in fact, in any of Classical Chinese up at least to the Hàn period. Thus, it is a rare (but not unique) isogloss that places EOC closer to post-Classical language than to the Classical epoch. Other quasi-synonyms have been excluded from comparison, such as 佳 (**krē* > *jiā*) 'beautiful, excellent' (met more rarely and generally in highly expressive contexts), 良 (**raŋ* > *liáng*) 'kind, good-spirited' (usually applied to human or animal nature rather than anything else), etc.

E. Excluded from analysis

E.1. 'lie': This (static) meaning is notoriously hard to separate from the closely related 'lie down, go to sleep' (dynamic) and 'sleep', not only in ancient texts, but in many modern dialec-

tal corpora as well (it is no wonder that it is very frequently omitted from various wordlists published in Chinese linguistic sources). The PTH equivalent is the recent innovation 躺 *tǎng*, of unclear origin; earlier literary sources mostly feature ambiguous data, with such quasi-synonyms as 寢 *qǐn* and 臥 *wò* translatable as 'go to sleep', 'lie down', or 'be sleeping' depending not only on the context, but on the translator's intuition as well. There is no formal ground in this case to speculate on possible lexical replacements in pre-PTH times.

Analysis

Having presented the data in its entirety, we can now proceed to the stage of analysis — a relatively brief one, since our only important task here is to calculate the number of replacements (or, more accurately, discrepancies, since we do not want to assume that each of the four analyzed stages was a direct linguistic descendant of the previous one). As could already be seen from the data, many cases in which such discrepancies were postulated are actually problematic and often derived from indirect evidence, particularly in the case of EOC vs. COC, where the attested corpus does not always allow us to resolve the issue of synonimity to complete satisfaction. For that reason, in the tables below I will discriminate between «certain» and «probable» replacements, where the former are clearly evident from sufficient textual evidence and the latter are based on insufficient and/or circumstantial evidence.

Additionally, in respect to the long transitional period from COC to MC it is useful to log the information on cases where a solid argument may be made for a lexical replacement already evident in Hàn-era literary texts (despite the lack of a separate wordlist for the Hàn period); such cases will be marked with a + sign next to the item in question.

	Certain replacements	Probable replacements
EOC → COC	'all', 'road', 'root', 'sleep'	'bite', 'breast (chest)', 'foot', 'good', 'man'
COC → MC	'belly', 'bite', 'cold', 'dog+', 'eat', 'eye', 'foot+', 'head+', 'say', 'sleep', 'tree', 'two', 'go', 'what'	'burn+', 'earth', 'smoke+', 'full+', 'neck+', 'that'
MC → PTH	'all', 'nail', 'give', 'hear', 'mouth', 'red', 'stand', 'sun', 'say'	'green', 'this', 'tooth', 'drink', 'egg'

Adding up both certain and probable replacements, we thus get the following picture:

1) 9 replacements over the approximately 400–500 year period separating EOC from COC;

2) 20 replacements over the approximately 1,200–1,400 year period separating COC from MC (of these, about a third may have taken place over the approximately 300-200 year period separating COC from Hàn-era Chinese, though this number is not fully confirmed);

3) 14 replacements over the approximately 800–1,000 year period separating MC from PTH;

4) altogether, 43 replacements from EOC to PTH (counting twice for those few items that have been replaced two times — 38 otherwise).

Quite importantly, none of the attested replacements can be reliably attributed to external borrowing; although for some of them (especially those that lack reliable Tibeto-Burman cognates) an original non-Chinese source is quite possible, the majority are first attested in texts with non-Swadesh meanings, so the replacements have to be judged as «internal». According to Sergei Starostin's revised methodology of glottochronological calculations, this means that we should expect the rates of change to be reasonably regular, without any periods of intensive speeding-up due to contact-induced processes of lexical intereference.

The results are not convincingly consistent with the division of the Swadesh wordlist into the less stable and more stable sub-sets as described, e.g., in Starostin 2010: although of all the

listed items, slightly less than half belong to the more stable sub-set ('nail', 'dog', 'drink', 'eat', 'egg', 'eye', 'foot', 'head', 'hear', 'mouth', 'smoke', 'sun', 'tooth', 'tree', 'two', 'what'), the proportion is still close to 50/50 and hardly significant. It does seem interesting that nearly all the reliable and potential replacements from EOC to COC fall into the less stable half of the wordlist, but whether this observation is historically important remains to be seen.

Conclusions

1. Taking Early Old Chinese as the starting point and Modern Chinese as the endpoint, we can claim, based on a mix of direct and indirect evidence from the text corpus (and some dictionary information), that *approximately* 60% of the Swadesh wordlist has been retained over 3,000 years of linguistic evolution. (The rounding-up of the percentage, rather than being an aesthetic concession, should hint at the possibility of errors in data analysis and occasional wrong conclusions based on insufficient data). This figure is not in direct contradiction either with the classic Swadesh formula (t = -ln(0.6) / 0.14 ≈ 3650 years) or with the revised Starostin formula ($t = \sqrt{}$ -ln(0.6) : 0.05×0.6 ≈ 4120 years), though it does obviously fit in better with Swadesh's assessment.

2. The individual replacement rates for the three checkpoints are as follows: ≈ 0.18 for EOC to COC, ≈ 0.14 for COC to MC, ≈ 0.14 for MC to PTH. Other than a slight increase in the first case (which could be explained by different factors, such as incorrect dating, errors in wordlist construction, or a significantly divergent dialectal base for EOC, meaning that the real time difference between it and COC should be higher), the results over different time periods seem to be impressively consistent — *and* in unexpectedly good agreement with Swadesh's classic *lambda* value of 0.14 for 1,000 years (rather than Sergei Starostin's 0.05 over the same period).

3. However, these figures may need slight corrections depending on whether we subscribe to the idea that the selected checkpoints are not necessarily in a straightforward ancestral relationship: for instance, the real time distance between MC and PTH may not be the 800–1,000 years that separate the text of the *Línjì lù* from today's colloquial Mandarin Chinese, but a period of as much as 1,000–1,400 years (to be more confident, one would have to conduct a very thorough and rigorous dialectal study of the text). In other words, observed *lambda* values might be slightly inflated (but only slightly: thinking of MC and PTH as two completely independent developments from COC or EOC is not supported by evidence).

4. If there is any circumstantial evidence for a one-time acceleration period, the best candidate would probably be the transition from COC to Hàn-era texts, where we witness, over a span of no more than 200 years, the replacement of such words as 'head', 'neck', 'foot', 'dog', and others. However, since the main dialect of Hàn-era texts is hardly a direct descendant of the Northern (Lǔ?) dialect that forms the basis for the COC list, it may be argued that at least some of these replacements could have happened earlier and are simply undetected due to lack of textual evidence from that dialect preceding the 3rd century BC (which brings us back to point 3).

5. It is particularly instructive to compare the acquired result with historically similar situations for other written languages, especially those already covered in the Global Lexicostatistical Database (Starostin ed. 2011–2019). Thus, for the Greek language (wordlists compiled and published by Alexei Kassian) we have a wordlist for the Ancient Attic dialect (4th century BC, largely based on the language of Plato), compared with Modern Demotic Greek: the number of lexical replacements is 39 (all of them internal, just like in Chinese), which gives a *lambda* value of ≈ 0.16, completely in line with our results for Chinese (unfortunately, no high quality wordlists for any forms of Byzantine Greek are as of now available in the GLD).

On the other hand, it is also true that comparison with another Indo-European situation, namely, Old Norse vs. Modern Icelandic, shows a different result: only 2 replacements ('eat', 'swim') over the approximately 700-800 years that separate the two stages, resulting in a lambda value of ≈ 0.025 (this result basically just repeats the observations already publicized in the well-known anti-glottochronological paper by Bergsland and Vogt, 1962). But what this shows, in my opinion, is not the simplistic «glottochronology does not work» conclusion that is drawn by many researchers, but rather that different rates of replacement may be triggered by different sociolinguistic situations — indeed, it may be argued that historically, the cases of Greek and Chinese have more in common with each other (large dialectal variety; co-existence of an archaic written language with evolving colloquial norms; active contact with neighboring languages) than either of them with Icelandic. Naturally, a full comparative analysis of these situations will only be possible after a detailed analysis of all the empirical evidence that may be gathered from other written languages across the globe (Indo-European, Semitic, Egyptian, etc.); hopefully, the present study takes a small step in the right direction.

References

Baxter, William H., Laurent Sagart. 2014. *Old Chinese: A New Reconstruction*. Oxford University Press.

Bergsland, Knut, Hans Vogt. 1962. On the Validity of Glottochronology. *Current Anthropology* 3: 115–153.

Dobson, W. A. C. H. 1968. *The Language of the Book of Songs*. University of Toronto Press.

Gurevich, Isabella S. 2001. *Lin-ji lu*. Saint-Petersburg: Peterburgskoje vosotokovedenije.

Hamed, Mahé Ben, Wang Feng. 2006. Stuck in the forest: Trees, networks and Chinese dialects. *Diachronica* 23(1): 29–60.

Kassian, Alexei, George Starostin, Anna Dybo, Vasily Chernov. 2010. The Swadesh wordlist: an attempt at semantic specification. *Journal of Language Relationship* 4: 46–89.

List, Johann-Mattis. 2015. Network perspectives on Chinese dialect history. *Bulletin of Chinese Linguistics* 8: 42–67.

List, Johann-Mattis. 2016. Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. *Journal of Language Evolution* 1(2): 119–136.

Pulleyblank, Edwin G. 1995. *Outline of Classical Chinese Grammar*. Vancouver: UBC Press.

Sawer, Michael. 1969. *Studies in Middle Chinese Grammar: the language of the early yeu luh*. PhD Thesis, Australian National University.

Schuessler, Axel. 1987. *A Dictionary of Early Zhōu Chinese*. Honolulu: University of Hawaii Press.

Starostin, George. 2010. Preliminary lexicostatistics as a basis for language classification: a new approach. *Journal of Language Relationship* 3: 79–117.

Starostin, George. 2013a. Lexicostatistics as a basis for language classification: increasing the pros, reducing the cons. In: H. Fangerau, H. Geisler, Th. Halling, W. Martin (eds.). *Classification and Evolution in Biology, Linguistics and the History of Science: Concepts — Methods — Visualization*: 125–146. Stuttgart: Franz Steiner Verlag.

Starostin, George. 2013b. K probleme dvux sobak v klassicheskom kitajskom jazyke: canis comestibilis vs. canis venaticus? In: N. P. Grintser et al. (eds.). *Institutionis Conditori: Ilje Sergeevichu Smirnovu. Orientalia et Classica, vol. L*: 253–267. Moscow: RSUH Publishers.

Starostin, George (ed.) 2011–2019. *The Global Lexicostatistical Database*. Moscow: Russian State University for the Humanities, & Santa Fe: Santa Fe Institute. Available online at: http://starling.rinet.ru/new100.

Starostin, Sergei A. 1989. *Rekonstruktsiia drevnekitaiskoi fonologicheskoi sistemy* [Reconstruction of the Old Chinese phonological system]. Moscow: Nauka. (In Russian).

Starostin, Sergei. 2000. Comparative-historical linguistics and lexicostatistics. In: Colin Renfrew, April McMahon, Larry Trask (eds.). *Time Depth in Historical Linguistics*: 223–259. McDonald Institute for Archaeological Research, Oxford Publishing Press.

Sturgeon, Donald (ed.). 2019. *Chinese Text Project*. Available online at: https://ctext.org.

Swadesh, Morris. 1952. Lexico-Statistic Dating of Prehistoric Ethnic Contacts: With Special Reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society* 96(4): 452–463.

Wu, Jianshe. 2011. The evolution of basic color terms in Chinese. *Journal of Chinese linguistics* 39(1): 76–122.

Г. С. Старостин. Китайская базисная лексика в диахронической перспективе и ее значимость для лексикостатистики и глоттохронологии

В статье сравниваются относительные скорости замены базисной лексики (представленной стандартным 100-словным списком Сводеша) на протяжении истории развития китайского языка, от раннедревнекитайского (представленного такими текстами, как *Книга песен*) к классическому древнекитайскому, позднему среднекитайскому (представленному языком памятника *Линьцзи лу*) и современному китайскому. В первой части статьи последовательно излагается методология составления списков; вторая посвящена детальному обсуждению всех обнаруженных лексических замен. В заключительной части показано, что в среднем скорость распада списка от одного периода к другому меняется незначительно, и что в целом результаты согласуются с классической «константой Сводеша» (0.14 замен за тысячу лет); более того, обнаруживается корреляция и с некоторыми другими аналогичными ситуациями, например, с историей греческого языка, хотя в отдельных случаях (исландский) такой корреляции не наблюдается. Можно надеяться, что дальнейшие исследования такого рода по лексической эволюции языков с длительной письменной историей позволят поместить полученные результаты в более широкий и значимый контекст.

Ключевые слова: история китайского языка, древнекитайский язык, среднекитайский язык, лексикостатистика, глоттохронология, базисная лексика.