Published in: Mahlberg, Michaela/ González-Díaz, Victorina/ Smith, Catherine (eds.): Proceedings of the Corpus Linguistics Conference 2009, Liverpool: University of Liverpool, 2010. 22pp.

Putting corpora into perspective
Rethinking synchronicity in corpus linguistics[1]

Cyril Belica, Holger Keibel, Marc Kupietz, Rainer Perkuhn
Institute for the German Language (IDS), Mannheim
corpuslinguistics@ids-mannheim.de

Marie Vachková
Charles University, Prague
vachkova@ff.cuni.cz

Abstract
Empirical synchronic language studies generally seek to investigate language phenomena for one point in time, even though this point in time is often not stated explicitly. Until today, surprisingly little research has addressed the implications of this time-dependency of synchronic research on the composition and analysis of data that are suitable for conducting such studies. Existing solutions and practices tend to be too general to meet the needs of all kinds of research questions. In this theoretical paper that is targeted at both corpus creators and corpus users, we propose to take a decidedly synchronic perspective on the relevant language data. Such a perspective may be realised either in terms of sampling criteria or in terms of analytical methods applied to the data. As a general approach for both realisations, we introduce and explore the *FReD strategy* (*Frequency Relevance Decay*) which models the relevance of language events from a synchronic perspective. This general strategy represents a whole family of synchronic perspectives that may be customised to meet the requirements imposed by the specific research questions and language domain under investigation.

1 Introduction
The most obvious prerequisite for conducting synchronic empirical studies is a synchronic corpus. But what does it mean for a corpus to be *synchronic* in the first place? Synchronicity is best described as a special aspect of representativeness which is itself a tricky core concept in corpus linguistics. Extrapolating observations from a corpus to a specific language domain – and this is the scientific interest and practice in most corpus-based work – is only justified when the corpus constitutes a sufficiently representative sample of this domain. However, because for most language domains the representativeness of a corpus cannot be evaluated in practice in a satisfactory way, the sampling of corpora usually seeks to approximate representativeness by intuitively estimating some qualitative and quantitative properties of the respective language domain and requiring the corpus to roughly display these properties, too, at least as far as time, budget and other practical constraints permit. A corpus that does display both kinds of properties is generally called *balanced* – or more precisely: balanced with respect to the estimated properties.

When corpus creators aim at composing a corpus that is balanced in this sense, they typically focus on the distribution and proportions of dimensions such as *mode* (spoken vs. written), *register* (fiction, news, academic, opinion, journal, etc.), *text type* (interview, comment, novel, short story, political speech, etc.), or *topic* (politics, economy, sports, science, etc.). The dimension of *time* (i.e., the time at which a corpus part was originally produced or published) generally receives much less attention, even when the corpus is intended to represent some contemporary language domain (e.g., contemporary British English) – as was the case for many popular electronic corpora, at the time of their creation. Most commonly,

representativeness with respect to time is approximated simply by including only language material that was produced in some prespecified time range (e.g., 1964-1994 for the written component of the British National Corpus). In some cases, the corpus is additionally required to be *balanced across time*, i.e., to contain roughly the same amount of data for each time slice (year, month, decade) in the given time period. One example is the *DWDS Kernkorpus* (core corpus) for the German language of the 20th century where decades are used as the unit of time slice (Geyken 2007).

This latter criterion may be described as *chronologically uniform sampling strategy* (short: *CUS strategy*). It implements a fairly straightforward model of time which, without doubt, is highly adequate, when composing a corpus to represent a language domain that is defined by a specific time period. For example, the time-related sampling strategy underlying the DWDS Kernkorpus is adequate for representing the German language of the 20th century, and any empirical research on this domain is well-advised to use as empirical basis a corpus build with the same or a similar strategy.[2] However, this uniform strategy implements a particular notion of *synchronicity* which may only apply to some language domains and research questions. In particular, so we will argue in section 2, it is generally not adequate for representing a language domain defined by a specific point in time such as *today* (rather than by a fairly large time period). To illustrate this again for the same example, the DWDS Kernkorpus may represent the German language of the 20th century very well, but it is unlikely to be a good sample of the German language as it was in the year 2000 – or, to overstate the point: as it was, say, on 31 December, 2000 at noon.

What is needed, is a more general approach offering adequate sampling strategies for all types of language domains and research questions. The main goal of this paper is twofold: (i) to outline such a general approach of sampling strategies which is grounded in epistemological, linguistic and psychological considerations; and (ii) to evaluate the empirical consequences of adopting these strategies. We would like to stress from the outset, that the time dimension is logically independent of any other dimensions (such as the ones listed above). In other words, any existing sampling strategies formulated for these other dimensions remain valid and may be combined with whatever sampling strategy we propose for time.

The remainder of this paper is structured as follows: In the next section, the general approach of sampling strategies is derived and described conceptually, before we provide a more formal definition in section 3. We explored the consequences of these strategies in several different ways, and the results of these explorations are summarised in section 4, while section 5 illustrates these overall findings on a few specific examples. In section 6, we discuss the properties of the proposed sampling strategies with respect to monitor corpora, which leads to some fundamental insights about these strategies in general. Implications of this work and possible future directions are discussed in the final section.

2 The fading relevance of language events

Assume we have some CUS-sampled corpus, and imagine three language phenomena whose chronological frequency distributions in this corpus look like those depicted in Fig. 1. The circles represent the different usage events of the respective phenomenon, and the vertical bars indicate the different time slices across which these events are distributed.

From the perspective of the CUS-sampled corpus, all three phenomena appear equally relevant, due to their identical overall frequencies. And they probably are, when the object of investigation is some language domain defined by the time period that is covered by this fictitious corpus. However, if one wishes to model the corresponding language domain language as it is *today* (e.g., contemporary British English), the three phenomena would probably be ascribed very different relevance for this domain, due to their markedly different

frequency distributions. The first phenomenon (depicted in the left-hand chart in Fig. 1) has continuously decreased in frequency so drastically that its relevance for the contemporary language domain is probably much lower than the other two phenomena. Likewise, the third phenomenon (right-hand chart) with its steep increase of frequency may be considered more relevant for today's language that the other two.

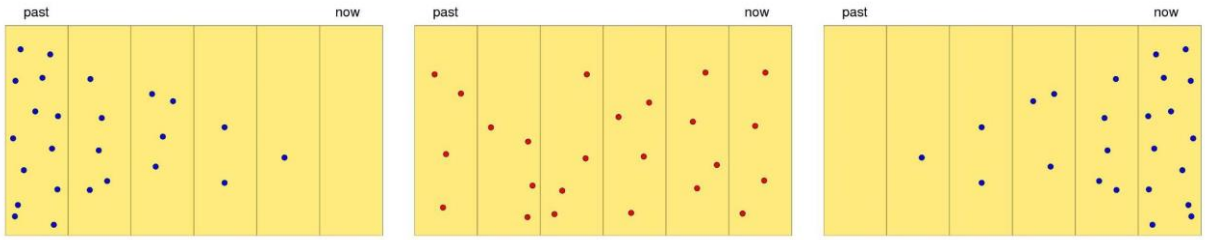

Figure 1. Three fictitious language phenomena and their distribution across time

Of course, one may argue that the second phenomenon (central chart), with its more evenly distributed occurrences, is less prone to fluctuations and therefore more deeply rooted in language than the other two. Its relevance for today's language should therefore be at least close to that of the third one. This is an important consideration which depends on the particular language domain to be represented. As we will point out later, such considerations may be incorporated in fine-tuning the general sampling strategy that we propose in this paper.

Nevertheless, the overall point is that because language change is a continuous process, more recent language productions generally tend to have more relevance for representing a contemporary language domain than older ones. An extreme conclusion might even go as far as stating that only the most recent language events – e.g., language productions of the last four weeks or some even smaller time window – provide a valid sample of the given language domain in its current state. But this would generally not lead to good statistical samples of the contemporary language domain: First, this approach would involve an *arbitrary discontinuity* in the concept of synchronicity – just like any CUS strategy does – in the sense that in four weeks from now, data produced today would *suddenly* cease to be considered *good data*, although they have been considered perfectly good data until just one day before. A second, more practical issue concerns the *data acquisition* as it would be fairly difficult to collect a sufficiently large sample of the desired kinds of language data that were produced in such a short period of time. Third, this approach would exclude a lot of potential variety in language use and capture only those words, constructions and other language phenomena that happened to have a sufficient number of occasions to become manifest in such a short time period. Thus, even if we had access to all sentences and utterances produced within such a short period of time, we would probably not obtain a good sample of the given language domain in its current state. In other words, corpus size is not the only relevant factor to ensure that especially rare events are well represented in the corpus – one also needs to incorporate data that stretch across time.

We would like to propose a less extreme conclusion which follows fairly straightforward – if not inevitably – from the above line of argument: namely that all past language events – as instantiations of the contemporary language domain – are relevant to some degree, but their specific degree of relevance is gradually fading over time. This conclusion leads in turn to the following, more general working hypothesis.

Frequency Relevance Decay (FReD):
If a phenomenon occurred at a certain frequency in a given time slice, the relevance of these occurrences for later points in time gradually decreases over time.

This hypothesised *FReD effect*, if correct, has strong implications on synchronic studies when *synchronicity* is defined in terms of a contemporary language domain: it urges us empirical linguists to take a *vanishing point perspective* on data of language use, implying a fading relevance of time slices with increasing "age", rather than the traditional *bird's eye perspective* where all data are weighted equally, irrespective of the time slice in which their were produced. This bird's eye perspective corresponds to what we described as CUS strategy. In this perspective, all data appear equally large, as if one would look at them from far above, in a timeless environment. By contrast, in a vanishing point perspective, the observer looks at the data from somewhere on the ground – viz. on the time-scale –, at the same level where the data themselves occurred.

These spatial analogies are meant to underscore the fundamental distinction between the two notions of synchronicity that different "synchronic" studies and projects may have in mind: synchronicity defined by a (fairly large) time interval vs. synchronicity defined by a point in time. Both notions are important, but which of them is valid for a particular study depends on the language domain to be investigated. As will become apparent in the next section, the notion of synchronicity-as-interval can be interpreted and modelled as an extreme instance of synchronicity-as-point. In the remainder of this paper, we therefore use the terms *synchronicity* and *synchronic* to refer to this latter notion.

In addition to the corpus-linguistic arguments stated above, there also is a cognitive argument for the hypothesised FReD effect. In cognitive linguistics, it is generally assumed that for individual speakers, the degree of entrenchment (i.e., routinisation) of a word, structure, etc. correlates with its experienced frequency. Likewise, to further extend this to a social argument, it is often, albeit implicitly, assumed for the language community that the degree of entrenchment (here referring to conventionalisation or typicality) of a word, structure etc. correlates with its observed frequency in an appropriate corpus. In both cases, if the respective frequency of a phenomenon gradually decreases over time, its degree of entrenchment is expected to gradually decrease, too, resulting in a gradual subjective or inter-subjective "forgetting", although the term *forgetting* should not be taken too literally here (cf. section 3). In this sense, the postulated FReD effect may be interpreted as counterpart of the process of entrenchment, as a form of *de-trenchment*. The bottom line of these cognitive and social arguments is that the FReD effect should be taken into account by any corpus-linguistic studies intended to investigate the current degree of entrenchment of a phenomenon.

Until today, provisions for the FReD effect probably have not been of crucial importance for corpus linguistics, at least with respect to written language: the amount of electronic text material available for corpus creators was constantly increasing over time, for various reasons. As a consequence, even when corpus creators did not pay much attention to the time dimension, the resulting corpora often happened to reflect a FReD sampling strategy. However, this situation may change in the near future, there may be a *ceiling effect* with respect to the amount of textual data available in each new time slice. As a consequence, the FReD effect will become increasingly important in the future – important for both corpus creators and corpus users. This paper offers no ready-made solutions off the shelf, its main purpose is to address the issue and outline possible approaches.

3 Formal modelling

3.1 FReD weighting function

The hypothesised FReD effect states that the present relevance of any past occurrences of some phenomenon gradually decreases over time. Although this is a continuous process, it is useful to approximate it by a discrete one. The formal model we propose therefore consists of a *FReD weighting function* which assigns each time slice a weighting factor that quantifies the relevance of language data produced in this time slice, from the perspective of the present. As a first guess, we chose a sigmoid curve as the general shape for the FReD weighting function as in Fig. 2.

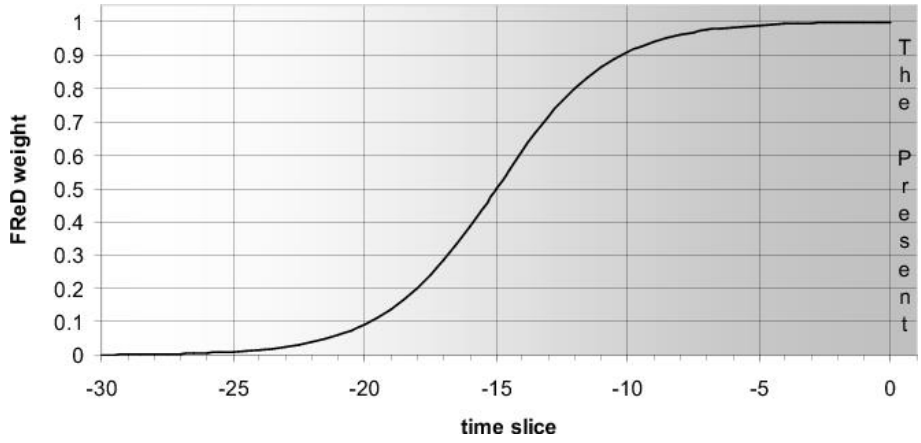

Figure 2. FReD weighting function: weighting of data as a function of time slice

The x-axis in Fig. 2 shows the discrete time slices (here representing arbitrary units of time), counted backwards from the present which represents the point in time defining the contemporary language domain under investigation. The y-axis gives the corresponding weighting factors. These weights are close to 1.0 for the most recent time slices, and close to 0.0 for the time slices that are old enough to be considered of little relevance for today's language.

The choice of this sigmoid function constitutes an assumption which will have to be justified by psychological experiments and other considerations. At first, the sigmoid shape might seem to conflict with what is known as the *forgetting curve* (Ebbinghaus 1885/1992) which displays an exponential decay instead of a sigmoid shape. Crucially, however, this forgetting curve and the FReD weighting function model two very different forms of forgetting such that their different shapes constitute no contradiction. Ebbinghaus' forgetting curve measures the individual forgetting of newly acquired explicit knowledge and memories, and this forgetting is generally assessed in terms of the decreasing success to retrieve bits of such knowledge. By contrast, the FReD weighting function models the relevance of past language events from a synchronic perspective. It is about the subjective and inter-subjective "forgetting" of implicit language knowledge (in the form of entrenched language routines or conventions, respectively), and this kind of "forgetting" may be assessed in terms of a decreasing degree of entrenchment. In other words, this kind of "forgetting" does not refer to losing (access to) explicit knowledge, but rather to language routines being used less routinely or conventions being considered less typical, respectively.

Formally, the FReD weighting function in Fig. 2 is implemented as a logistic function $f(t)$ defined by formula (1).

$$(1)\quad f(t) = \frac{f_0}{f_0 + (1 - f_0)\cdot e^{k\cdot(t - t_0)}} \quad \text{with} \quad k = \frac{\ln(1/f_0 - 1)}{T_H}$$

3.2 Parameters

The FReD weighting function as defined above assigns the relevance weight $f(t)$ to time slice t. It provides three parameters by which this function may be varied. Thus, one actually gets a whole family of FReD functions.

First, the *"now" parameter* t_0 specifies the fixed point in time – or time slice – that is defined as the *present* by the given language domain. This is the reference point from which the relevance weights of past time slices are assessed. By default, this reference point will be the present now of real-life time. In the particular function underlying Fig. 2, this "now" parameter is set to time slice 0.

The second parameter T_H controls what one might term the *"half-life"* of the decreasing relevance: namely, the number of time slices it takes for the full relevance weight (1.0) to decrease down to 0.5 which in this logistic model takes place at the function's inflection point. In the particular function underlying Fig. 2, the "half-life" parameter is set to 15 time slices such that the inflection point is located in time slice -15. Note that the term "half-life" is not fully adequate here as it usually refers to exponential decay, but for lack of a more accurate label we use it anyway.

Finally, the parameter $f_0 = f(t_0)$ sets the maximal weight – i.e., the weight ascribed to time slice t_0. This maximal weight will usually be a value just below 1.0. While the other two parameters implement aspects of the FReD model that may directly relate to the respective language domain that one wishes to study, relating this third one is less intuitive. It effectively controls the slope of the FReD curve, in particular the *slope at the inflection point* which can be determined from f_0 (and the "half-life") by a non-linear monotonic transformation. In consequence, the greater the parameter f_0, the steeper is the curve around the inflection point and at the same time the flatter towards either end (presuming the "half-life is held constant; cf. Fig. 3). Thus, instead of prespecifying the parameter f_0, one might equivalently choose the slope at the inflection point. We therefore do not distinguish these two quantities and refer to both as the *slope parameter*.

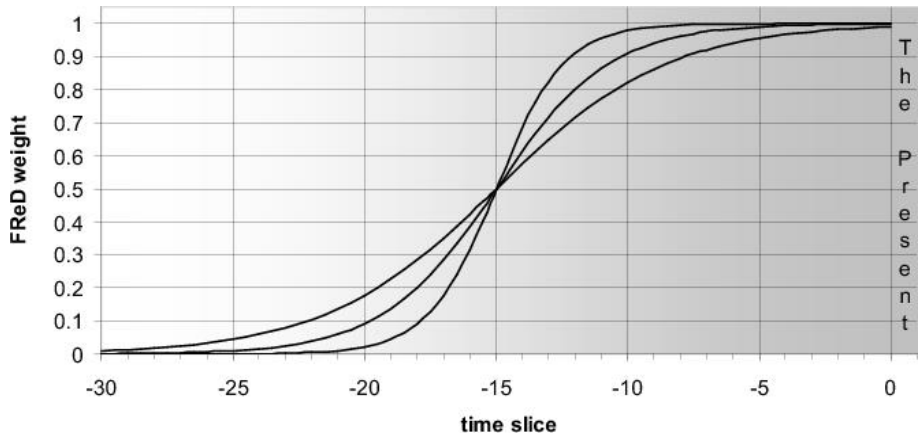

Figure 3. Effect of varying the slope parameter, everything else being equal

Interestingly, by gradually increasing the slope parameter f_0 is gradually increased towards 1.0 – without ever reaching 1.0 – the FReD weighting function approximates a step function (cf. Fig. 4) which corresponds to the traditional bird's eye perspective, i.e., the CUS strategy with the discontinuity at time slice -15. At the other extreme, when varying the slope in the other direction and proceed like this indefinitely, the weighting function approximates a linear function (also shown in Fig. 4). The linear function and the step function thus can be seen as extreme instances of the general FReD weighting function which in turn offers a controlled transition between these two extremes. In other words, the approach proposed here can be conceived of as an extension of existing strategies.

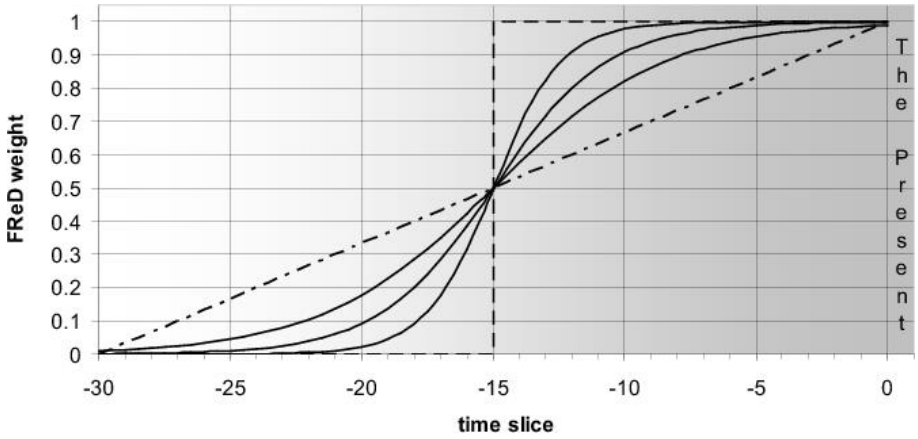

Figure 4. Approximating two extreme weighting functions, step function (dashed line) and linear function (dash-dotted line), by FReD functions (solid lines)

Fig. 5 displays different FReD functions obtained by varying the "half-life", while keeping slope and "now" parameter constant.

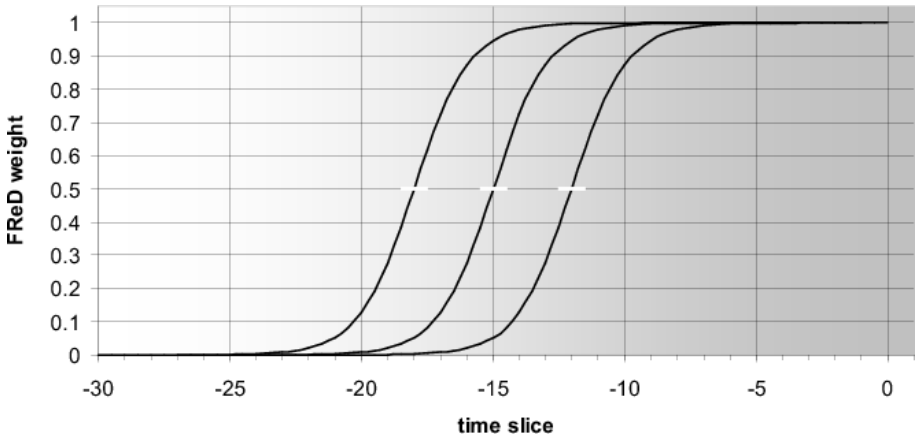

Figure 5. Effect of varying the "half-life" parameter, everything else being equal.

In all FReD functions plotted so far, the "now" parameter was set to the present "now" in real-life time, here presented as 0. However, if for instance we build a synchronic corpus from the perspective of today, then freeze this corpus and wait for, say, seven time slices (e.g., years), the same corpus could then be described by a FReD weighting function as in Fig. 6. Here, however, another option is more intriguing: this particular function can be used to build such a corpus directly in seven years from today – this would then amount to asking in retrospect: what was synchronic seven years ago?

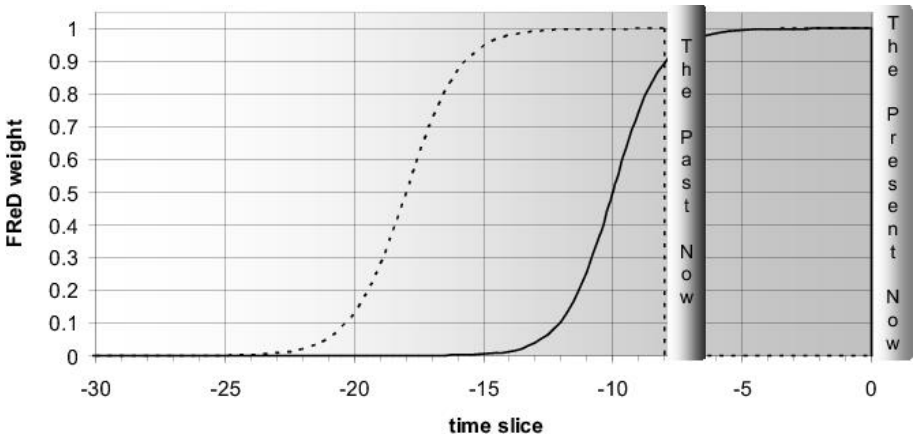

Figure 6. Effect of varying the "now" parameter, everything else being equal.

We believe that with its three parameters, the general FReD weighting function defined by formula (1) offers enough flexibility to suit the needs of a broad range of synchronic studies. Ideally, each study would work with particular parameter settings that are fine-tuned to several factors, most importantly to (i) the language domain under investigation, (ii) to the specific research questions pursued by the study, and (iii) to the number and size of selected time slices that seem optimal for these research questions. These three factors, most of all the first one, represent a specific notion of synchronicity, and ultimately, each study has to define for itself a notion of synchronicity that describes best what the researchers have in mind as their object of investigation. For example, a study analysing contemporary German teenage slang is likely to opt for a very different notion of synchronicity than another study interested in academic writing where language change presumably proceeds less quickly and drastically. The FReD weighting function does not constitute an answer to this question, but with its customisable parameters it enables researchers to implement their specific notion of synchronicity in an appropriate way.

3.3 Weighting of what

So far, we have interpreted the FReD weighting function in a rather vague manner as quantifying for each time slice the relevance of language data produced in this time slice, from the perspective of what is considered as the *present*. This raises the question how exactly the relevance weights of the different time slices – or rather, of the data produced in them – may be implemented in practice. In other words, how exactly may a FReD strategy be incorporated by synchronic studies?

There are at least two possibilities which may be characterised as *adjusting frequencies* and *corpus sampling*. The former one concerns not so much corpus composition but corpus analysis. In this case, one starts from a corpus that is sampled with a chronologically uniform strategy, and for any quantitative analysis of any language phenomenon one uses weighted

corpus frequencies wherever normally raw frequencies would be used. Formally, the weighted absolute frequency of a phenomenon in the overall corpus is determined by simply multiplying its frequency in each time slice with the corresponding FReD weight and adding up the individual products across time slices. For instance, if a fictitious phenomenon occurred twice in a time slice with weight .7 and once in a later time slice with weight .85, the overall weighted frequency would be 2.25 (=0.7·2+0.85·1). In other words, with this first type of implementation, the FReD weighting function defines *frequency weights* for time slices. In this case, one does not actually build a new type of corpus, but only looks at a traditional corpus in a new way.

In theory, adjusting frequencies is how the FReD strategy should ideally be implemented. Unfortunately, however, it raises a number of practical problems, as it would require the adaptation of any corpus-analytical method that is based on quantitative data, e.g., collocation extraction. Especially the fact that the adjusted frequencies are generally non-integer values (as in the above example) might impose additional requirements on the statistical techniques, and some techniques might not be applicable to adjusted frequencies at all.

A much simpler approach is the second type of implementation which can be interpreted as an approximation of the first one. Here, a new kind of corpus is built, by extracting a random sample of whole texts (a so-called "virtual corpus") from a text repository that serves as super-sample or "primordial sample" (Kupietz & Keibel 2009: 56). In this case, the FReD weighting function is used to define the relative *sampling sizes* (e.g., number of texts) for the different time slices. Such a *FReD sampling strategy* is conceptually simple and fairly easy to accomplish. It constitutes an approximation of the adjusting frequencies approach quite literally in that any absolute frequencies derived from a FReD-sampled corpus may be interpreted as an approximation of the corresponding adjusted frequency in the first approach. However, this interpretation involves the implicit assumption that the distribution of text sizes in the underlying primordial sample is roughly constant over time. Otherwise, a smaller unit of sampling should be chosen (e.g., sampling by paragraph or sentence instead of text).

The primary disadvantage of this second possibility is that it affects the statistical robustness, in at least two ways: First the sampled corpus will generally be smaller than a corresponding corpus used for the adjusting frequencies approach. Second and closely related, extracting whole texts involves the rounding of floating point numbers (viz., the FReD weights) to integers (viz., the number of texts extracted in a time slice), and these rounding effects concern especially the older time slices with lower FReD weights. In other words, the rounding effects are particularly relevant for language phenomena that have occurred most frequently in the older time slices. The magnitude of the rounding effects may be cushioned by using smaller sampling units such as paragraphs or sentences instead of texts, as suggested above.

3.4 Summary

In sum, employing a FReD strategy for conducting a synchronic study involves a sequence of scientific decisions. As the very first of these decisions, it is crucial to be fairly clear about the language domain to be investigated, and likewise about the specific research questions that one wishes to ask about this language domain. This has consequences on what counts as the right type, amount and variability of corpus data, and the availability of such data, of course, constrains the kinds of language domains and research questions that may reasonably be pursued. But more importantly, this first decision should always involve stating as explicitly as possible the particular notion of synchronicity underlying the study (cf. 3.2).

A second decision concerns the number and size of relevant time slices. The optimal values depend on the language domain and research question, but as a rule of thumb the size of time slices should be large enough for a decent amount of data to be available for each of them, and at the same time small enough that potential changes of the phenomena of interest are likely to

stretch across multiple time slices. The optimal number of time slices is closely intertwined with customising the FReD function, and may therefore also be considered a part of the next decision. In any case, the number of time slices should fit the FReD weighting function – or vice versa – in the sense that the FReD weight of the oldest time slice that is still included is virtually zero.

As the third decision, the parameters (e.g., "now", "half-life", slope) of the FReD weighting function have to be set and fine-tuned to the first two decisions. Next, one has to decide about the type of implementation of the FReD strategy (adjusting frequencies vs. corpus sampling), and if corpus sampling is selected, there are two additional modelling decisions to be made: one concerns the granularity of the sampling unit (text, paragraph, sentence, etc.), the other the desired corpus size which, of course, is limited by the maximum amount of data available for a single time slice.

4 Explorations

In order to evaluate the consequences of adopting a FReD strategy, we conducted a range of explorations for specific language phenomena. This is a way of testing the plausibility of the general FReD weighting function as a measure of relevance of language events from a synchronic perspective. The ultimate goal of these explorations therefore was to gain confidence that a FReD strategy helps to achieve the same adequacy with respect to time, as does a balanced sampling with respect to mode, register or text type.

In the explorations, we took advantage of DEREKO (the Mannheim German Reference Corpus; Kupietz & Keibel 2009) and extracted from it all issues of the daily newspaper *die tageszeitung* (*taz*) in the period from 1989 through 2008. The motivation for using only a single newspaper was to ensure a high degree of homogeneity for the explorations. These *taz* data served as our primordial sample from which we derived two virtual corpora, defined by two different sampling models: (i) a FReD sampling strategy, resulting in a "forgetting" corpus, and (ii) a chronologically uniform sampling strategy (i.e., CUS), resulting in a non-forgetting corpus. For the FReD strategy, the "now" parameter was set to 2008, the "half-life" to 9.5 time slices (each time slice representing one year), and the slope parameter f_0 to 0.999. For the CUS strategy, the weight for each time slice was set to 0.5 such that the overall sizes of both corpora are roughly identical (approx. 116 million words each). Fig. 7 illustrates how the two sampling functions look like.

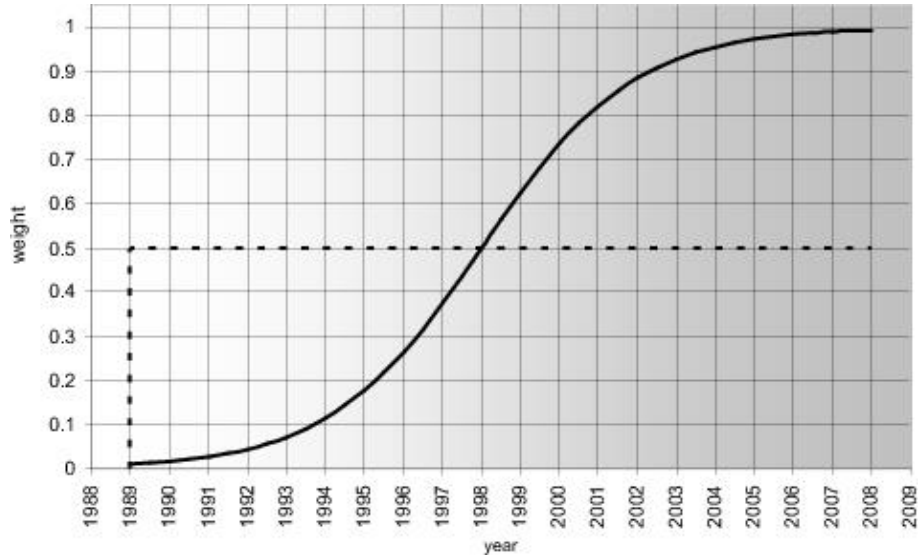

Figure 7. Sampling functions used for FReD sampling (solid line) and CUS sampling (dashed line).

These two corpora were used to explore the consequences of using a FReD strategy in contrast to a traditional CUS strategy. To this end, we focused on phenomena with a skewed frequency distribution across time because for non-changing phenomena, both sampling strategies would be indistinguishable. It should be stressed, however, that a FReD strategy is meant to model the relevance of all language phenomena – stable and unstable ones – as seen from a synchronic perspective defined by the "now" parameter.

We automatically extracted candidate lists for chronologically unstable phenomena at different levels: word frequencies, collocations, similar collocation profiles, meaning potential of words, and near-synonyms. We then hand-picked several candidates from each list and evaluated – on the basis of native speaker competence – whether the FReD-sampled corpus offers a more realistic synchronic view of the respective phenomenon than does the CUS-sampled corpus.

In the next section, we provide only a few examples of these explorations. To summarise our overall findings for all types of explorations, most of the automatically extracted candidates turned out be of little interest in terms of language change, as the main factor causing their changing corpus properties across time seemed to pertain to real-world events (e.g., Germany's re-unification, the war on Iraq, etc.), the German spelling reform, the introduction of the euro, and proper names. It appears that many genuine language-internal changes are more subtle and in some cases even rather slow processes such that they do not stand out as much in terms of salient changes in a newspaper corpus which mainly captures public discourse on language-external events. Nevertheless, in our candidate lists, we still did find a number of instances that apparently do relate to genuine changes in language use.

Among these instances, we encountered no counter-intuitive cases – that is, no cases where the observations in the CUS-sampled corpus would appear to be closer to our intuitions than those in the FReD-sampled corpus. In other words, these explorations do provide supportive evidence for the claim that the general FReD strategy is a plausible way of modelling a language domain from a synchronic perspective, but ultimately, stronger evidence from other sources (such as psychological studies) is needed to demonstrate this plausibility.

5 Examples

As stated in the previous section, all our explorations were contrastive, in that we analysed any example phenomena for the FReD-based corpus relative to the CUS-sampled corpus. When interpreting the following examples, it is therefore important to always think of the CUS-based results as the observations one would have considered *synchronic*, if there were no FReD strategy.

5.1 Word frequencies

The first set of examples concerns word frequencies. Consider the following two words: *ernstlich* (English: serious, seriously) and *frau* (feminine variant of the personal pronoun *man,* English: *one*). This latter word has originally become popular because the pronoun *man* is not generally perceived as referring to both genders equally as it is derived from *Mann* (English: *man*) and is homophonic with it. Fig. 8 displays the frequency distributions of both words across the 20 time slices for the CUS-sampled corpus. In this corpus, the usage frequency of both words has decreased fairly continuously over the past 20 years. From a synchronic view, both words are probably still considered a part of the language but less relevant than 20 years ago. This is reflected more strongly in the FReD-sampled corpus where the overall frequencies of *frau* (799) and *ernstlich* (149) are lower than the corresponding frequencies for the uniform corpus (1154 and 181, respectively). The overall relevance ascribed to these two words is thus lower for the FReD strategy. The interaction between the frequency development of a word and its relevance according to the FReD strategy is explored more closely in the next section on monitor corpora.

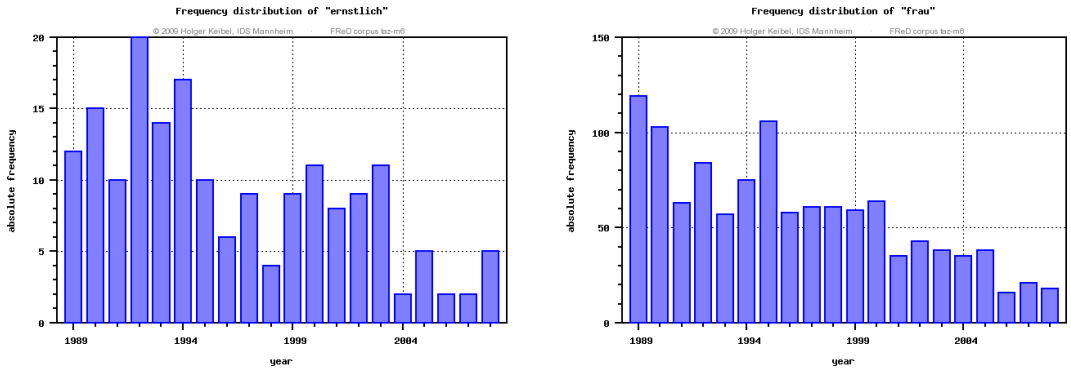

Figure 8. Frequency development of *ernstlich* (left) and *frau* (right) in the CUS-sampled corpus

A closer inspection of the occurrences of *ernstlich* in both corpora revealed more subtle and unexpected changes about this word's use beyond the mere decrease of frequency. Inflected forms of *ernstlich* are usually attributive adjectives, while the non-inflected form itself is used as a predicative adjective or as an adverbial. Interestingly, there is a tendency towards the non-inflected use in the FReD-sampled corpus, when contrasted with the CUS-sampled corpus. It is likely that this finding is a result of recent changes in the usage preferences of this word, i.e., genuine changes in the language itself. This finding was entirely unexpected, and it suggests that, if the FReD-sampled corpus is indeed a more adequate empirical basis for a synchronic study, the non-inflected use of this word is underestimated by the CUS-sampled corpus and should be ascribed greater relevance. Conversely, if we had available independent evidence about the (increasing) preference of speakers/writers for the non-inflected form of *ernstlich*, this would then support the FReD strategy for implementing a synchronic perspective.

5.2 Collocations

Another set of candidate phenomena were collocational patterns that changed over time. As one such example, consider the noun *Machenschaft* which is almost exclusively used in the plural form *Machenschaften* (English: doings, schemes/scheming, intrigues, illegal activities). We determined its collocates in both corpora, making use of the collocation algorithm by Belica (1995; cf. Keibel & Belica 2007). Many collocates of *Machenschaften* are found for both corpora, most salient are *kriminellen, kriminelle, dunkle, dunklen, illegalen* (English: criminal, dark, illegal). Some collocates, however, are specific to the CUS-sampled corpus, e.g. *illegaler, skandalösen, betrügerischer, dubiose, unseriösen, unlauteren, bösen* (English: illegal, scandalous, dubious, shady, fraudulent, dishonest), while others are only observed for the FReD-sampled corpus, e.g. *mafiösen*, *menschlichen*, *bizarren*, *rechte*, *gewisse* (English: mafia-like, human, bizarr, right-wing, certain). Thus, as expected, the FReD strategy does have consequences on the cohesivesness of word combinations.

Recall that relative to the CUS-sampled corpus, the FReD-sampled corpus emphasises the more recent time slices. Therefore, one possible explanation for these collocational differences between both corpora is that the meaning aspects underlying the collocations specific to the CUS-sampled corpus may have become a part of the core meaning of the word *Machenschaften* itself such that speakers today generally do not feel a need to explicitly mention these aspects any more, it is taken for granted that all *Machenschaften* are illegal by nature. At the same time, speakers apparently tend to highlight the specific kind of illegal activities that they talk about, especially by whom are they carried out (by the Mafia or mafia-like groups, by right-wing groups, etc.). By the intuition of competent speakers of German, this explanation seems plausible, and this plausibility in turn provides support for the plausibility of the FReD strategy which, however, needs to be verified psychologically.

5.3 Similar collocation profiles

Based on the same flexible notion of collocation, we derived large collocation profiles for more than 200,000 lemmas. Each such profile consists of the full spectrum of significant collocations around the respective lemma. If the collocation profile of a lemma is interpreted as representing the lemma's usage preferences, lemmas that tend to be in used similar ways are expected to have similar profiles, and vice versa. In much of our previous work, we have used a formal measure of similarity between collocation profiles (Belica 2001-2007) to explore the similarity structure between words, in various different ways (e.g., Keibel & Belica 2007, Vachková & Belica 2009, Belica in press; Belica et al. under review). These explorations not only verified the general plausibility of the similarity measure but also confirmed the prediction about the correlation between a word's usage properties and its collocation profile. Moreover, they gave rise to the development and evaluation of several analytical methods that exploit the similarity structure between words (cf. the same papers).

We applied these methods also for the present explorations on the consequences of adopting a FReD strategy, and in this and the following subsections, we provide several examples of these analyses. For the first example analysis, consider the adjective *krass* (English: extreme, blatant, terrible). We determined for each corpus the list of words whose collocation profiles were most similar to that of *krass,* the top portions of these lists are given below.

(2a) CUS-sampled corpus:
kraß, fundamental, eklatant, schwerwiegend, gravierend, auffällig, grotesk, fett, offenkundig, fatal, grob, wiegen, daneben, Koalitionsvertrag, tragisch, Geschlecht, eindrucksvoll, Vordergrund, erschrecken, absurd

(2b) FReD-sampled corpus:
kraß, eklatant, geil, auffällig, mies, gravierend, fett, schwerwiegend, scheißen, grob, sauer, ausgeprägt, unglaubwürdig, witzig, cool, exemplarisch, scheiße, blöde, blöd, fatal

Each of these two lists may be interpreted in isolation, leading to insights about the meaning potential of *krass* for the respective corpus. Comparing both lists in turn leads to insights about changes in the word's meaning potential. This comparison reveals striking differences: for the FReD-sampled corpus, *krass* is similar to many words that are used mainly in colloquial language whereas there is only one such item (*fett*) in the other list, and even this is not necessarily an indicator of colloquiality. It is known for spoken German that *krass* has assumed a colloquial reading (which would translate to cool!, gosh!, or wicked!), and our current observations suggest that this reading has also entered written German. Again, this aspect would be missed or underestimated by a chronologically uniform sampling strategy.

5.4 Meaning potential

Self-organising maps (SOMs; Kohonen) have proven a useful data-driven methodology for visualizing and studying the complex semantic structure of a word (e.g., Keibel & Belica 2007, Vachková & Belica 2009, Belica in press; Belica et al. under review). In the SOM of a word *x*, other words that are similar to *x* (in terms of similar collocation profiles) are used to study the semantic potential of *x*. The SOM presents these other words on a grid such that proximity on the grid reflects similarity in terms of use. Visually scanning such an SOM will generally prompt competent speakers to identifying several regions or clusters of words relating to specific usage aspects of the given word *x*, or more precisely: to specific global, language-external contexts in which *x* is used. Like this, the methodology is capable of guiding a linguist in explicating their implicit knowledge of the complex semantic properties of a given word (cf. the same sources as above).

For the purposes of the present study, we looked at a range of interesting words and generated for each of them two SOMs, one for each of the two corpora, and had a competent speaker manually annotate the SOMs for putative global contexts. Fig. 9 shows the resulting annotated SOMs for *ernstlich* (English: serious, seriously). Several regions can be identified, most of them are identical for both sampling models. However, in the SOM based on the FReD-sampled corpus, there is one region that apparently refers to usage aspects of *ernstlich* pertaining to environmental and social problems, whereas the SOM for the CUS-sampled corpus shows no traces of such a region. One possible explanation for this observation is that the topic of environmental and social problems may have received more attention, recently, an alternative explanation could be that *ernstlich,* which overall has become less frequent over the course of the past 20 years, at least in the *taz* data (cf. 5.1), may have survived in the niche of these topics, or even replaced some other word in these contexts. Whereas the former explanation mainly concerns changes in the language-external world, the latter refers to some genuine phenomenon of language change.

Likewise, in the SOM for the CUS-sampled corpus, we observed a region pertaining to what may be labelled “morals, habits and customs”, and there is no corresponding region in the other SOM. Again, this may be caused by language-external factors – i.e., these topics may, over time, have become less important in public discourse – or by actual language change, if another word has replaced *ernstlich* in its function in these topics.

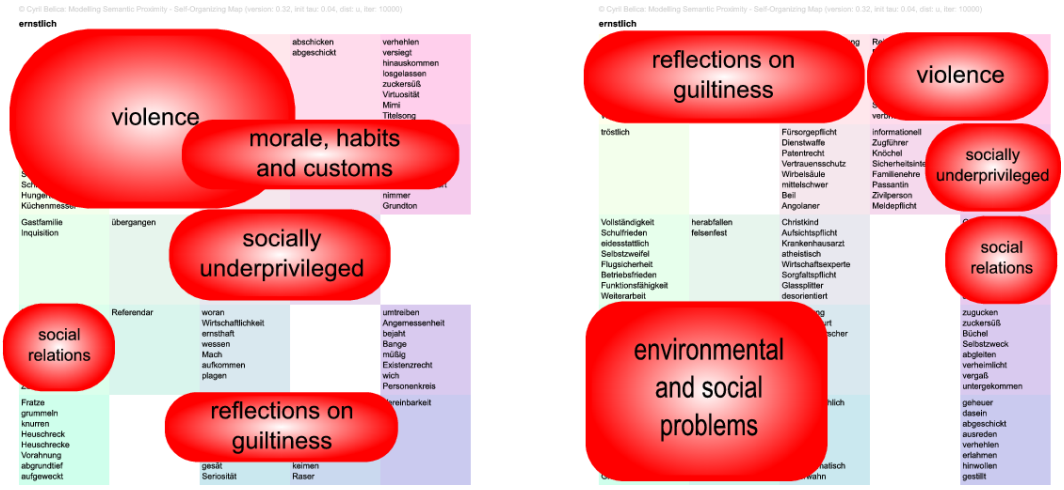

Figure 9. Annotated SOMs of *ernstlich* for the CUS-sampled corpus (left) and the FReD-sampled corpus (right)

In an analogous SOM analysis for *krass*, a region stands out in the FReD-based SOM that strongly relates to colloquiality, whereas there are very few and only scattered traces of colloquiality in the corresponding SOM for the uniform corpus. A straightforward interpretation of this latter SOM would therefore probably not prompt a competent speaker to assign the aspect of colloquiality to any region. This confirms the earlier observation for *krass* in subsection 5.3.

Figure 10. SOMs of *krass* for the CUS-sampled corpus (left) and the FReD-sampled corpus (right)

5.5 Near-synonyms

The same basic SOM methodology also serves the study of the relation between any two near-synonyms *x* and *y*. The only change is that now the SOM visualises the complex similarity structure between the items that are among the most similar words for either *x* or *y* (or both, for that matter). The grid elements in the resulting *contrastive SOM* are colour-marked such that the items in yellowish elements tend to be more similar to *x*, whereas the items in the more reddish grid elements lean more towards *y*. Orange grid elements display no preference for

either *x* or *y*, but are on average equally similar to both. Orange areas thus point to usage aspects shared by both *x* and *y*, whereas yellowish and reddish areas point to aspects that are unique to *x* or *y*, respectively (cf. the work referred to in 5.4, for a more detailed description of this extended methodology).

Fig. 11 shows two contrastive SOMs for the near-synonym pair *holen* (English: take, fetch) vs. *nehmen* (take, grab), on SOM for each corpus. In both SOMs there is a yellowish area which apparently relates to a sports context. This indicates that *holen* has a by far greater preference to be used in a sports context, and very typical instantiations of this preference are probably statements about an athlete or a team winning points, medals or championships (*Punkte/Medaillen/Titel holen*).

However, this preference of *holen* in contrast to *nehmen* is apparently constant across the two corpora. As in the previous subsection, what is of most interest in this present study, are qualitative differences between the two SOMs. Most prominently, the contrastive SOM for the CUS-sampled corpus contains a distinct region pointing to usage aspects that are unique to *nehmen* which, however, are difficult to interpret and require further explorations (concordances etc.). There is no corresponding area in the FReD-based SOM indicating that the degree of synonymy between the two words may have increased recently. This change, too, requires explanation, and without further investigations, the current example therefore does not contribute any supporting nor conflicting evidence about the adequacy of the general FReD strategy.

Figure 11. Contrastive SOMs of *holen* vs. *nehmen* for the CUS-sampled corpus (left) and the FReD-sampled corpus (right)

A second example of this type of analysis concerns the pair of synonyms *kriegen* vs. *bekommen* (both: get, receive). The two contrastive SOMs are shown in Fig. 12. The most striking observation is that while the SOM for the CUS-sampled corpus reflects a fairly high degree of overall synonymy between the two words, the FReD-based SOM contains a distinct region for *bekommen* which appears to relate to social and financial benefits, but also to other global contexts which require more detailed analyses. The same global conctexts also stand out in the other SOM, but there, they are associated with both words almost equally. Thus, it seems that the degree of synonymy of *kriegen* and *bekommen* has decreased recently, and again, this finding motivates further investigations.

Figure 12. Contrastive SOMs of *kriegen* vs. *bekommen* for the CUS-sampled corpus (left) and the FReD-sampled corpus (right)

6 FReD strategy and monitor corpora

So far, we have discussed and explored the FReD strategy only with respect to static corpora. The situation gets conceptually more complex, however, when this strategy is realised for monitor corpora.[3] A *synchronic monitor corpus* can be defined as a sequence of static FReD-sampled corpora, with a moving "now" parameter and the other parameters held constant. Importantly, such a corpus is defined by two variable points in time which are logically independent: (i) the point in time at which any given language event took place, and (ii) the moving "now" parameter defining the point in time from which the language events are looked at. There are interesting and nontrivial interactions between these two variable points in time, as will become apparent in the following example.

Consider the word *Jahrtausend* (English: millennium). Obviously, there were good reasons and plenty of occasions for this word to be used more frequently around the recent turn of the millennium. Of course, this increased frequency is caused by language-external factors and therefore not very interesting in terms of language change. However, the point of this example, in contrast to those in the previous section, is only to illustrate the relation between the two variable points in time in synchronic monitor corpora.

Fig. 13 shows the frequency distribution of *Jahrtausend* across time, for two monitor corpora: one composed by a chronologically uniform sampling strategy, the other by a FReD sampling strategy. In this example, both strategies only consider data in a sliding time window embracing five years, which would probably be too small for serious synchronic studies, but suffices to make the point. For the sake of clarity, the two frequency distributions are plotted in comparison to a *reference distribution* which simply consists of the raw relative frequencies observed for each year in isolation.

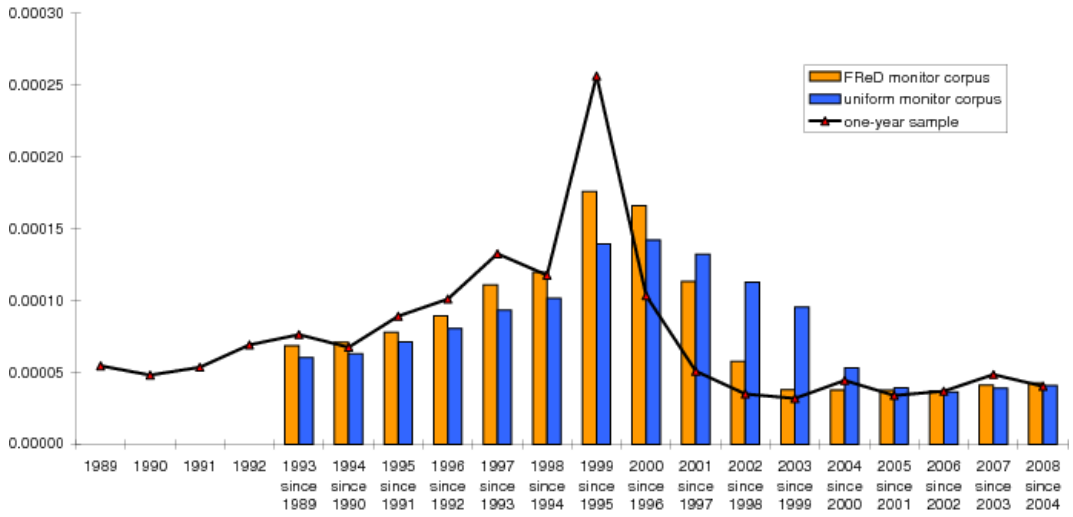

Figure 13. Year-wise relative frequency of *Jahrtausend* for a uniform, CUS-sampled corpus and for a FReD-sampled monitor corpus, compared to the observed relative frequency per year.

The frequencies for both monitor corpora lag behind the pattern of the reference distribution (cf. Fig. 13). This is because both monitor corpora have a memory and have fully forgotten a language event only after some time (in this example: five years) has passed. But forgetting takes place very differently for the two corpora: for the FReD-sampled monitor corpus, forgetting is a continuous process stretching across a longer period of time. In the CUS-sampled corpus, by contrast, forgetting is no process but an isolated event: there is initially no forgetting at all before it happens in just a blink of an eye. It is therefore not surprising to observe in Fig. 13 that the delay relative to the reference distribution is more pronounced for the CUS-sampled monitor corpus. In other words, the FReD-sampled monitor corpus responds more quickly to changing frequencies, and this concerns both increasing and decreasing frequencies.

This last observation illustrates an important aspect of the general FReD strategy: namely, that it is sensitive to the interaction of *frequency* and *recency.* It corresponds to an assumption according to which more frequent occurrences of some phenomenon that have taken place longer ago may have a similar influence on this phenomenon's present degree of entrenchment , as do less frequent occurrences that have taken place fairly recently. This assumption is a direct consequence of the hypothesised FReD effect – it is in fact equivalent to it.

There are a problematic and a welcome side effect of this simultaneous sensitivity to frequency and recency which concerns temporary changes of frequency. For instance, imagine a phenomenon whose frequency has increased only temporarily before resuming its previous frequency level (like the word *Jahrtausend* above). If the time period of this increase is fairly short it is unlikely that it corresponds to a genuine and lasting instance of language change – in most such cases, the temporary increase is caused by language-external events and the public discourse about them. Compared to a CUS strategy, a FReD strategy responds more quickly and even more strongly to such a temporary increase, but afterwards, it stabilises much faster (cf. Fig.13). Therefore, the FReD strategy may be said to overemphasise apparent changes that took place fairly recently – a side effect which is clearly not desirable and should be controlled for. On the other hand, the CUS strategy overemphasises this temporary increase for a much longer period of time – in the above example, it does not forget about it at all until five years later, whereas the FReD strategy starts “forgetting” about this temporary increase fairly quickly.

In sum, the FReD strategy cushions the effects of most language-external events, as long as they did not happen very recently – relative to the “now” parameter – and the discourse about them did not last too long. This does not only concern changes in observed frequency, but also

changes with respect to cohesion strength, word meaning, and so forth. For instance, among the collocates of *Machenschaften* for the CUS-sampled corpus (cf. section 5.2), there are obvious traces of a political scandal (the so-called *Waterkant-Gate*) that took place in 1987, whereas no such traces are observed for the FReD-sampled corpus. In this latter corpus, however, *Machenschaften* collocates with words that unambiguously relate to more recent political scandals intensely debated in public discourse, and these collocates are in turn not observed for the CUS-sampled corpus.

7 Discussion

7.1 Summary

The general FReD strategy proposed in this paper is an attempt to model the relevance of previous occurrences of some language phenomenon with respect to its degree of entrenchment at a given later point in time. This relevance is modelled as an interaction of frequency and recency which constitutes an assumption that needs to be tested in psycholinguistic experiments. If valid, the assumption predicts that for a language phenomenon whose frequency of use is decreasing over time, its degree of entrenchment is also gradually decreasing, with this latter decrease lagging behind the former one. Psychologically, the decreasing degree of entrenchment may be interpreted as a process of "forgetting", not in the sense of a reduced success of retrieving some explicit knowledge, but rather in the sense of implicit language knowledge being less routinely used, and correspondingly, of a language convention being commonly perceived as less typical.

The theoretical considerations and empirical explorations at different language levels that were presented in this paper suggest that the general FReD strategy is a plausible way of realising a synchronic perspective. A conclusive demonstration that it is not only plausible but also adequate requires, again, further supporting evidence from experimental studies. Provided that such evidence will be established in the future, the bottom line of this work is to trust the frequencies derived by a FReD strategy: phenomena that appear rare or archaic or new or cohesive asf. by a FReD-sampled corpus (or in terms of FReD-adjusted frequencies) may be confidently treated as such. One exception concerns temporary changes in the observable data which are mainly caused by fairly recent language-external events (cf. section 6).

The general FReD strategy may implement a whole range of different synchronic perspectives, and a specific one of these perspective is chosen by estimating the values of several – in our case three – independent parameters. In practice, these parameters should be fine-tuned to the specific language domain and research questions of interest, in such a way that the intended extension of the term "synchronicity" is adequately operationalised. Any specific FReD strategy may be implemented in two ways (adjusting frequencies vs. corpus sampling), and irrespective of this type of implementation, it is applicable for realising a synchronic view on language that is either static (i.e., defined by a fixed "now" parameter) or dynamic (with a moving "now" parameter).

7.2 Future directions

In order to make the general ideas outlined in the paper more fruitful, possible future work should involve experimental studies to test the sigmoid FReD curve as a general model of the present degree of entrenchment of language phenomena. Ideally, such studies would operationalise the notion of (individual and collective) entrenchment – e.g., in terms of reaction times – and measure its development as a function of time and frequency. The central research question will be, whether and to which extent a decreasing frequency of use in fact results in reduced entrenchment.

To extend this line of though one step further, later studies may find for some language domain that the processes of language change – in the form of changing degrees of entrenchment – has accelerated over time. If this is indeed the case, the FReD weighting function will need to be modified such that the "half-life" and slope parameters are set dependent on the "now" parameter, especially with respect to synchronic monitor corpora.

One more practical problem of the general FReD strategy is that, in its present form, it is not readily applicable within any specific corpus-linguistic studies. What is needed here is a set of guidelines for estimating the parameters of the weighting function, and these guidelines in turn should be motivated by independent empirical work.

Another type of follow-up study with practical implications would empirically assess the reliability of the FReD implementation by corpus sampling as an approximation of the theoretically more preferable other type of implementation (adjusted frequencies).

A fundamental shortcoming of the FReD approach is that it models a highly complex issue – viz. the relevance or entrenchment of language events from a synchronic perspective – in terms of a fairly simple model which so far only incorporates time and frequency of use. Obviously, there are several other potentially influential factors that should be taken into account: e.g., production vs. reception, age of acquisition (for subjective entrenchment), dispersion, political events, ethical issues, emotional charge, periodic events, demographic development (as affecting what counts as language change), etc. Here, "political events" and "ethical issues" are not intended to refer to language-external events as such, but rather to the fact that such events and issues may potentially establish conventions about certain words or phrases being politically incorrect or preferable which may in turn affect speakers' usage preferences. In any case, it is an open research question for most of these and other factors whether they have a significant impact on the degree of entrenchment of language phenomena. Any study trying to pursue this question will be faced with the challenge that it is unclear how these factors may be assessed in practice, even if this may be possible only by means of some fairly indirect and approximate operationalisation.

Among the potentially influential factors listed above, one might be more easy to incorporate in the FReD approach, namely the corpus-linguistic concept of *dispersion* which is motivated as follows. The corpus-based frequency of some phenomenon may, by itself, be very misleading and should be interpreted relative to how evenly or unevenly the occurrences of this phenomenon are distributed across the various corpus parts (cf. Gries 2008). This degree of *dispersion* is therefore likely to be an influential factor with respect to entrenchment. However, in our view, it is important to treat the concept of dispersion as independent from the time dimension. First, the hypothesised time effects on entrenchment (viz. the FReD effect) cannot be accessed by a global dispersion measure even if it is specialised to *dispersion across time*, for it will only quantify the overall degree to which the frequency distribution across time deviates from a uniform distribution, but not in which ways it deviates (e.g., increasing or decreasing over time). Therefore, time effects on entrenchment are modelled in a much better way by the FReD strategy in its present form. Second, if later, some measure of dispersion is to be included in this strategy, it should be insensitive to dispersion across time, because otherwise the influence of time on the degree of entrenchment would effectively be assessed twice. In other words, if the notion of dispersion is to be incorporated in a model of entrenchment, it should be based on corpus parts that are not defined by time.

7.3 An optimistic future scenario

We conclude the paper with outlining the kind of possible future scenario in which the FReD strategy would be most useful. In this scenario, one would have available a virtually unlimited quantity and variety of language material for each potentially relevant time slice, that is, an extremely large and well-stratified primordial sample (cf. 3.3). In such a scenario, it would then

be possible to compose, for any given language domain and research question, a corresponding synchronic corpus of a predefined overall size *N*. One would merely have to define a specific FReD function (by choosing the model parameters) and to extract from the given primordial sample a specialised subsample whose relative sizes for the different time slices are prescribed by the FReD weights.

To our knowledge, this optimistic future scenario revolving around a sufficiently large and stratified primordial sample is far from being accomplished for any language. As a long-term goal, however, it corresponds to the design and ambition underlying the corpus archive DeReKo (cf. section 4) which currently comprises roughly 3.75 billion text words and has an average growth rate of approximately 300 million words per year.

Notes

1. The authors wish to thank Sophie Hennig (IDS Mannheim) for helping with the example analyses.
2. This is in fact the case for the ongoing DWDS project (http://www.dwds.de/) which aims at building a large dictionary for the German language of the 20th century on the basis of the DWDS Kernkorpus.
3. The authors like to thank an anonymous reviewer of an earlier abstract submitted to the Corpus Linguistics conference for emphasising this aspect of the FReD strategy.

References

Belica, C. (1995). *Statistische Kollokationsanalyse und Clustering. Korpuslinguistische Analysemethode.* © 1995 Institut für Deutsche Sprache, Mannheim.

Belica, C. (2001-2007). *Kookkurrenzdatenbank CCDB. Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs.* © 2001-2007 Institut für Deutsche Sprache, Mannheim.

Belica, C. (in press): "Semantische Nähe als Ähnlichkeit von Kookkurrenzprofilen". In: *Korpusinstrumente in Lehre und Forschung / Corpora: strumenti per la didattica e la ricerca / Corpus Tools in Teaching and Research*. Bozen: alpha beta piccadilly Verlag. Available at: http://corpora.ids-mannheim.de/SemProx.pdf (accessed: 10 Dec 2008)

Belica, C., H. Keibel, M. Kupietz and R. Perkuhn (under review). "An empiricist's view of the ontology of lexical-semantic relations". In P. Storjohann (ed.) *Lexical semantic relations from theoretical and practical perspectives*.

Ebbinghaus, H. (1992). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie.* (Neue, unveränd. und ungekürzte Ausgabe nach der 1. Auflage 1885). Darmstadt: Wissenschaftliche Buchgesellschaft.

Geyken, A. (2007). "The DWDS corpus: A reference corpus for the German language of the twentieth century". In C. Fellbaum (ed.) *Idioms and collocations: Corpus-based linguistic and lexicographic studies*. London: Continuum, 23-40.

Gries, St. Th. (2008). "Dispersions and adjusted frequencies in corpora". *International Journal of Corpus Linguistics*, 13 (4), 403–437

Keibel, H. and C. Belica (2007). "CCDB: A corpus-linguistic research and development workbench". *Proceedings of the 4th Corpus Linguistics Conference*, Birmingham.

Available at: http://corpus.bham.ac.uk/corplingproceedings07/paper/134_Paper.pdf (accessed: 28 January 2009)

Kohonen, T. (1984). *Self-organization and associative memory*. Berlin: Springer.

Kupietz, M. and H. Keibel (2009). "The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research". *Working Papers in Corpus-based Linguistics and Language Education*, No. 3. Tokyo: Tokyo University of Foreign Studies (TUFS), 53-59. Available at: http://cblle.tufs.ac.jp/assets/files/publications/working_papers_03/section/053-059.pdf (accessed: 12 June 2009)

Vachková, M. and C. Belica (2009): "Self-organizing lexical feature maps. Semiotic interpretation and possible application in lexicography". IJGLSA 13 (2), 223-260. [Interdisciplinary Journal for Germanic Linguistics and Semiotic Analysis, Berkeley: University of California Press]
Available at: http://corpora.ids-mannheim.de/IJGLSA.pdf (accessed: 8 Sep 2009)