Investigation into the linguistic category membership of the Finnish planning particle *tota*

Abstract

Even though hesitations (e.g., *um/uh*) were historically perceived as involuntary non-linguistic items (e.g., Maclay & Osgood, 1959), more recently, a number of scholars have suggested that hesitations can behave like (a) lexical items (e.g., Clark & Fox Tree, 2002) and (b) at least in some contexts and with some functions as grammatical items like suffixes/clitics (Kirjavainen, Crible & Beeching, 2022; Tottie, 2017). The current study contributes to this body of work and presents two spoken language corpus analyses (frequency analysis; network analysis) investigating the nature of the Finnish planning particle *'tota'.* Our results suggest that '*tota'* is more similar to grammatical items than lexical items.

Investigation into the linguistic category membership of the Finnish planning particle *tota*

1. Introduction

In addition to words, morphemes, and syntactic structures, spoken language (and sometimes, written language, Tottie, 2017) contains items, often referred to as filled pauses or fillers (e.g., *um, uh*), that are relatively high frequency vocalizations (Bortfield, et al., 2001; Fox Tree, 1995; Shriberg, 1994) but do not fit in neatly into the aforementioned conventional linguistic categories. Early studies often viewed these items as paralinguistic disfluency markers related to language production difficulties (e.g., giving speakers extra time to plan their utterance or find a word) (e.g., Brennan & Schober, 2001; Corley & Stewart, 2008; Levelt, 1983; Maclay & Osgood, 1959) and categorized them as non-linguistic items with little semantic meaning, even though some systematic uses were associated with filled pauses already early on (e.g., Boomer & Dittman, 1962; Maclay & Osgood, 1959). More recently, often focusing on English filled pauses (*um*, *uh*), researchers have suggested that these are linguistic items that have self-repairing, hesitation and planning functions, but importantly also pragmatic functions (e.g. Clark & Fox Tree, 2001, Götz, 2013; Tottie, 2011) and that some filled pauses could be categorised as grammatical or lexical items (e.g., Clark & Fox Tree, 2002; Kosmala & Crible, 2022; Kirjavainen, et al., 2022; Schneider, 2014; Tottie, 2011; 2015; 2017). The current study contributes to this body of research and investigates a less studied language, Finnish, to see if the *planning particle* (Hakulinen, et al., 2004: Table 129, §792) '*tota*' that behaves similarly to the English filled pause '*um'* in that it has hesitation and pragmatic functions (see e.g. Etelämäki & Jaakkola's (2009) work on the usage of *tota*) could be seen as a grammatical or lexical item, rather than a less meaningful paralinguistic vocalization. In this paper, we will refer to *tota* as a *planning particle*, but do not intend to indicate that *tota* would not have pragmatic functions (i.e., that its function would only relate to speech planning) or that it would not show any functional overlap with some discourse markers e.g., *no*(*niin*) 'well'.

We will analyse a spoken Finnish corpus and report two studies. To establish that the planning particle *tota* has a different function from Finnish filled pauses (*ee* and *öö*) that have a strong hesitation function and that *tota* is thus not purely a *symptom* (Clark & Fox Tree, 2002) of production difficulties, in Study 1 we conduct an analysis of the surface frequency of the word tokens that immediately follow *tota* and *ee* and *öö* in a given speaker turn. As low frequency words are more likely to create word finding difficulties and thus hesitation, this analysis will inform us if *tota* has a hesitancy function similar to that of *ee* and *öö*. In addition, Study 1 assesses the distribution of the words that immediately follow *tota* and *ee* and *öö* at a part-of-speech level. This is to see if *tota* and *ee* and *öö* behave in the same way. According to the idea of distributed semantics (see Firth, 1957's claim "you shall know a word by the company it keeps", see also, Harris, 1954), words get their meaning in contacts: a word's meaning is defined in terms of (a) other words it tends to occur with, and (b) the words that tend to occur with words in (a) (Mitchell, 2019). That is, if *tota* and *ee* and *öö* occur with different (types of) words, they are also likely to have different meanings.

To further investigate the nature of *tota*, in Study 2 we conduct a network analysis in which we analyse the co-usage of *tota* and (a) Finnish filled pauses with hesitation function, (b) discourse particles with self-repair function, (c) politeness lexemes, (d) grammatical politeness markers, (e) planning lexemes and (f) repair lexemes. The aim of the network analysis is to see which items form usage clusters and to which of them the filled pause *tota* belongs. These two studies together can inform us as to whether *tota* is a hesitation marker, a lexical item or a grammatical item.

1.1 Disfluency markers

Filled pauses or other similar items are vocalizations that occur frequently in spontaneous speech (e.g., Bortfeld, et al., 2001; Fox Tree, 1995; Shriberg, 1994). These vocalizations are language dependent and can have one or more than one realization - in English *um* and *uh* are typically used, in Russian it is *eto* and in Finnish for example *tota*, *eiku* and *öö*.

Even though listeners often view hesitations as markers of disfluency (e.g., Fox Tree, 2001, 2002; Reynolds & Paivio, 1968), filled pauses have been found to facilitate language comprehension. A number of studies suggest that filled pauses help reference resolution and reduce the difficulty associated with new (e.g., Arnold, et al., 2004; Barr & Seyfeddinipur, 2010), complex (e.g., Watanabe, et al., 2008), low frequency (Bosker, et al., 2014) and low-predictability (Corley, et al., 2007) words. In addition to giving information about the referent type, filled pauses are also informative for the listener in indicating the level of certainty or truthfulness. Loy, Rohde, and Corley (2017, 2018) studied how speech disfluencies affect listeners' perception of a message by exposing participants to true or false information that was vs. was not preceded by a filled pause. The authors found that the listeners were sensitive to filled pauses produced by the experimenters and used them as pragmatic cues to assess as to whether the utterances were true or false – the presence of a filled pause indicating that what the speaker had said immediately after the pause was not true.

In language production, filled pauses have several functions. They are more likely to precede content words than function words (Maclay & Osgood, 1959) and infrequent and complex words than frequent or simple words (e.g., Beattie & Butterworth, 1979; Schnadt & Corley, 2006); they are also commonly produced at phrase or clause boundaries (Boomer & Dittman, 1962; Maclay & Osgood, 1959) indicating that filled pauses can be produced to give speakers time to plan what they want to say. This suggests filled pauses can have a hesitation function, for example, when the speaker is searching for a word (see example (1)). They often occur in restarts (e.g., Brennan & Schober, 2001), indicating an additional repair function (2). They have also been identified to occur as pragmatic markers indicating, for example, the level of certainty and politeness (e.g., Fox Tree, 2001; Smith & Clark, 1993; Brennan & Williams, 1995; Rendle-Short, 2004; Swerts, 1998; Swerts & Krahmer, 2005) (3) and being involved in turn-taking (e.g, Clark & Fox Tree, 2002; Kjellmer, 2003; Kosmala & Crible, 2022; Schegloff, 2010; Tottie, 2016) (4).[1]

(1) *What's the new lecturer's name?* ***Uh****, Felicity.*

(2) *What is the new lecturer's name? Fion…* ***uh*** *Felicity.*

(3) ***Um*** *would you be able to help me with this?*

(4) *I found the lecture pretty helpful* ***um***

These functions can be associated with different filled pauses available. For example, in English *uh* is more commonly used than *um* in disfluent contexts (e.g.,

[1] Examples 1-4 were created for the purpose of demonstrating the different functions.

during lexical identification and repair context) while pragmatic functions are more commonly expressed by *um* (e.g., Fox Tree, 2001; Rendle-Short, 2004; Swerts, 1998). The differential usage of the two filled pauses is further supported by studies investigating filled pause use in speakers with autism spectrum disorder (ASD) - while ASD speakers' usage of *uh* is similar to neurotypical speakers, their use of *um* is not (e.g., Irvine, et al., 2016; McGregor & Hadden, 2020). Because ASD patients typically have a pragmatic deficit, these studies indicate that in addition to hesitation, *um* also has pragmatic functions, while *uh* may be more closely related to disfluency.

Finnish has a wider range of items whose use corresponds to that of *um* and *uh* in English. The planning particle *tota (*or *tuota)*[2] can be used in hesitation contexts (5) but filled pauses *öö, ee* and *hh* also readily appear in these contexts (5) (Etelämäki & Jaakkola, 2009; Hakulinen, et al., 2004: §861; Laakso & Lehtola, 2003; Penttilä, et al., 2019). In addition to marking hesitation, *tota* also has pragmatic functions expressing politeness and the level of certainty (6), and turn-taking (7). It is not generally used in repair contexts, in which a number of different items, for example *eiku* 'no but', *nii(n)ku(n)* 'kind of, like', *siis* 'that is' or *tai* 'or' are used (Haakana & Visapää 2014; Hakulinen, et al., 2004: 822; Sorjonen & Laakso 2005). Thus, while *öö, ee* and *hh* and *eiku, niinku, siis* and *tai* correspond in their function to the English *uh, tota* has a similar function to the English *um*.

(5) E: hh niin ja katoj L sai siit rahaa ku se ol niinkö prosenttipalkal siäl hh
'uh and see L got money from it because he was there like on a commission'
P: ni hän sai provissiot siit

[2] In colloquial Finnish, the diphthong *uo* in the form *tuota* typically undergoes a reduction, in particular in spoken contexts, which leads to the form *tota* (Hakulinen, et al., 2004).

’so he got commission for it’

E: nii siit tuli se kymmene miljoonaa se sit niinkö nykyrahassa niin niin useampi euroo se sai

’yeah that added up to ten million that like in nowadays’ money that is that is he made quite a few euros’

P: joo ne o iha järjettömii ne ase ase **tota** hinna ko sillon ko se ol se **öyy** se **tota** L ol niit niit **öö** kranaatihettimii esittelemäs tua reserviupseerikerhol ni ku s niist hinnoist ku se sanos et tommonen Tampelan kevyt kranaatiheiti ni se om piäne auto hinnas

’yeah those weapon **um** prices are crazy because when he was he **uh** he **um** L was demonstrating those those **uh** grenade launchers at the reserve officers’ club and about the prices he said that one Tampella small grenade launcher that it is roughly the same price as a small car’

E: no se o älytön hinta

’well that’s a crazy price’

(Sapu115, 327-331)

(6) M: mthh niin niin semmosta ku mehäm mietittiin sitä Hannelle lähtööh

‘So that we have thought of going to Hanne’s’

S: ni

‘yeah’

M: hh niin **tota** a mä aattelin että hh että ä kyl se varmaa ijos päästäisköhän me molemmat sinne yöks et pitäskö mun kysyyh Hanneltah

‘So **um** I was thinking that uh that if I should ask Hanne if both of us could stay at hers overnight’

S : *niim mitä*

‘like what?’

M: et ku niinku lähtisik sää yöks sinne jos me e lähettäs sunnuntaina mthh ku

’I mean would you come along to stay at hers overnight if we go on Sunday uh like’ (SG113, 209–213)

(7) N: onko tota m hh onko ollu nyt hiljasta vai kiirettä vuosi

’has it um has it been quiet or busy this year’

T: no kyllä se kyl se hiljasempaa on on ollu nytte **tota**

‘yes it has been more quiet recently **um**’

N: mä menin vähän lankaan siinä ku mä kattelin vaan viime tammikuuta mut sillohan tuli ne vheronphalauthukset nhii hh mä

’it got me when I looked at the last January but then there were those tax returns so I’

T: hh heh

‘yep’ (SG108, 321–322)

The planning particle *tota* derives from the determiner/demonstrative pronoun *t(u)ota* 'that' (e.g., *Saanko vähän **t(u)ota** kakkua*? ‘Can I have some of **that** cake’; *Saanko vähän **t(u)ota**?* ‘Can have some of **that**?’) possibly via the process of grammaticalisation/pragmaticalisation. When *tota* is used as a planning particle its original (pronoun) meaning is bleached and it is hence semantically more similar to Finnish filled pauses *öö, ee,* and *hh* than to the determiner/demonstrative pronoun ‘t(u)ota’. However, research on disfluencies in Finnish aphasic patients suggests that some pronouns (e.g., *toi* ‘that’) might also be repeated in word finding contexts (Laakso & Lehtola, 2003) thus indicating that the pronoun and planning particle *tota* are likely to form a continuum from the more filled pause-like use to the more pronoun-like use in these contexts.

1.2 The linguistic status of filled pauses

Even though we know a fair amount about the usage of filled pauses and other similar items, the linguistic status of filled pauses (e.g., *um*, *uh* in English) has been debated for decades. Clark and Fox Tree (2002) refer to the different viewpoints as: (1) *filler-as-symptom*, (2) *filler-as-signal*, and (3) *filler-as-word*. Filler-as-symptom viewpoint assumes that filled pauses are non-linguistic vocalizations or noise that are produced with little overt intention and have very little meaning attached to them (e.g., Maclay & Osgood, 1959; Levelt, 1983). The proponents of the filler-as-signal viewpoint assume that while filled pauses might have some linguistic functions, they could or should not be assigned to any traditional word classes (e.g., Schnadt & Corley, 2006; Corley & Stewart, 2008). Others argue that they should be viewed as fillers-as-words for the following reasons:

- filled pauses have typical positions within utterances and their syntactic positions and form affect interpretation (e.g., Fox Tree, 2001; Rendle-Short, 2004; Swerts, 1998)
- they form chunks with words (e.g., Clark & Fox Tree, 2002; Crible, et al., 2017; Kirjavainen, et al., 2022; Schneider, 2014; Tottie, 2017)
- they can be planned for (e.g., they can occur in offline processes such as writing, thus demonstrating deliberate and planned inclusion of the filled pause, Tottie, 2017)
- they can be reduced or eliminated from language production depending on the context, thus demonstrating a level of control (Clark & Fox Tree, 2002).

That is, filled pauses should be seen as linguistic items that can be categorised into a word class (e.g., Clark & Fox Tree, 2002; Tottie, 2011; 2015; 2017).

A further debate around the nature of filled pauses involves the question: If filled pauses are fillers-as-words what kind of linguistic items are they? That is, which linguistic category do they belong to? Some have suggested that filled pauses could be seen as interjections or other similarly behaving stand-alone lexical items that do not integrate with the other linguistic units within an utterance (e.g., Biber, et al., 1999; Norrick, 2015) or as interjections that can be attached as a clitic-like item to words (Clark & Fox Tree, 2002). Others argue that filled pauses could have a more integrated function in sentences and could be seen, e.g., as adverbs, particularly in writing (Tottie, 2017). Filled pauses can also be used like clitics or suffixes that are attached to co-ordinating conjunctions and other function words (Clark & Fox Tree, 2002; Schneider, 2014; Tottie, 2017) and to verbs (Kirjavainen, et al., 2022) and form chunks such as *and-uh* and *said-um.*

The clitic-like filled pauses might be functioning as planning fillers (Tottie, 2017) or as pragmatic fillers (Kirjavainen, et al., 2022). Kirjavainen et al. (2022) investigated if, similarly to known linguistic items, distributional frequency patterns affected the representation of filled pauses used in pragmatic contexts in adult English speakers. Based on their corpus analysis data (study 1), the authors asked their participants in experimental contexts to repeat sentences in which the location of filled pauses and the words that occurred immediately before and after the filled pause was manipulated (study 3), measuring (a) repetition accuracy and (b) when the participants did not produce a verbatim repetition, the error types produced. They found the participants' repetition accuracy was affected by these manipulations, in particular when the filled pause followed its collocate (*said-um*). Furthermore, their analysis of errors of the repetition attempts suggested that when the participants heard *um* in an infrequent pre-verb position with its collocate *said* (*Mary* ***um said*** *Edinburgh was beautiful*) they typically moved the *um* to the frequent post-verb position with that collocate (*Mary* ***said um*** *Edinburgh was beautiful*), while this effect was significantly lower for a similar word (*thought*) that co-occurred with *um* infrequently (*Mary* ***um thought*** *Edinburgh was beautiful*). The authors argue that because filled pauses show frequency effects they behave similarly to fillers-as-words, and they interpret the error patterns (where the post-word position yielded stronger results than pre-word position) to suggest that filled pauses with pragmatic functions (*um*) might be processed similarly to grammatical items, such as clitics or suffixes.

The present paper will contribute to this body of research and investigates if the Finnish disfluency items, in particular the planning particle *tota* that has hesitation and pragmatic functions can be seen as filler-as-word, and whether it behaves more like a paralinguistic hesitation marker, a grammatical item or a lexical item. Even though some research exists for Finnish disfluency markers (e.g., Jansson-Verkasalo, et al., 2021; Penttilä & Korpijaakko-Huuhka, 2019; Penttilä, Korpijaakko-Huuhka & Bona, 2022), as far as we are aware, the question as to what kinds of items Finnish disfluency markers are has not been investigated before.

1.3 The present study

The current paper aims to answer the following two questions:

1. Is *tota*, a Finnish planning particle that can have a hesitation but also pragmatic functions (e.g., expressing politeness and/or the level of certainty,

and being involved in turn-taking) a *filler-as-word*, and different from *filler-as-symptom* (i.e., more paralinguistic hesitation)?

2. What linguistic category does *tota* belong to?

We will conduct two analyses on a spoken adult corpus. In Study 1, we will analyse the surface token frequency of words following *tota* and filled pauses *ee* and *öö*. Both are used in planning and word finding contexts in Finnish (see e.g., Laakso & Lehtola, 2003; Penttilä, et al., 2019), and also analyse what types of words *tota* vs. *ee* and *öö* typically precede. This study will inform us as to whether or not planning particle *tota*, which has seen in the literature primarily as a hesitation and disfluency item (Hakulinen, et al., 2004: Table 129 §792; §861; but see Etelämäki & Jaakkola, 2009), behaves similarly to other planning items (like *ee* and *öö)*. In Study 2 we will further investigate the status of *tota* by conducting a network analysis (Newman, 2010) which can be used to see if *tota* clusters together with (a) filled pauses with a hesitation function, (b) discourse particles with repair function, (c) grammatical politeness markers, (d) lexical politeness markers (e) hesitation and planning lexemes and/or (f) a repair lexeme. Study 2 informs us about the linguistic category membership of *tota*.

2. Study 1 – Distribution of *tota* vs. *ee* and *öö*

To investigate if the distribution of *tota* vs. *ee* and *öö* in speech is the same, we conducted three analyses as we detail below. We analysed

(1) the surface frequency of the words that immediately followed *tota* and *ee* and *öö*. Given that low frequency words are typically preceded by hesitation (e.g., Bosker, et al., 2014), if we find that *ee* and *öö* occur more commonly before low frequency items than *tota*, it would indicate that *tota* does not have as

strong a hesitation function as *ee* and *öö.* If we find no such difference, it would indicate both are used as hesitation markers.

(2) the parts of speech that followed *tota* and *ee* and *öö*. For this, we analysed the frequency with which *tota* vs. *ee* and *öö* are followed by different word types (noun, pronoun, verb, adjective, numeral, interjection, conjunction, adverb, none, i.e., end of turn). The main point for this analysis was to see if *tota* and *ee and öö* occur with the same kinds of parts of speech. Differences between the distribution of *tota* and *ee* and *öö* would indicate that their usage is not the same and subsequently that they are different types of items. Also, this analysis can inform us specifically about hesitant usage. Some syntactic positions (e.g., clause boundaries) are major planning points in speech and result in more hesitation (Goldman-Eisler, 1968; Levelt, 1983; Maclay and Osgood, 1959; Shriberg, 1994). If we find that *ee* and *öö* more commonly precede conjunctions than *tota*, it would indicate that *tota* does not have as strong a hesitation function as *ee* and *öö.* Given that ends of turns are unlikely to be planning points of speech, if we find that *tota* occurs at the ends of turns more commonly than *ee* and *öö*, it would indicate that *tota* does not have as strong a hesitation function as *ee* and *öö*. In the same vein, if we compare nouns and verbs and their use following *tota* vs. *ee* and *öö*, we expect that (on average) verbs, not nouns, should be associated with a more cognitive load while processing sentences (cf. to agrammatic speakers with aphasia who tend to omit verbs significantly more than nouns, see Menn and Obler, 1990). Therefore, if *ee* and *öö* are more likely to precede verbs than *tota*, we can assume that *ee* and *öö* have a stronger hesitation function than *tota*. For the adverbs, we expect that *tota* is more likely to precede adverbs than *ee* and *öö*

are. The reason is that *tota* is often used in crystalised phrases such as *tota noin* and *tota niiku* (e.g, Etelämäki & Jaakkola, 2009) in which the second constituent is semantically a very light adverb. Regarding the distribution of adjectives and numerals, we do not have a priori hypotheses.

(3) the types of nouns that follow *tota* vs. *ee* and *öö*. To analyse hesitation from non-hesitant uses further, we also coded each noun that followed *tota* vs. *ee* and *öö* as either a 'basic' noun (e.g., *chair*), a proper noun (e.g., *Coca-Cola* or *John*), a compound noun (e.g., *child care*), or a noun pronounced in English instead of Finnish (and which was not a conventionally accepted loan word; e.g., if a person said *teibl* instead of *pöytä* 'table'). We focused on the latter three noun types for the following reasons. Research suggests that proper names are associated with difficulties in word retrieval and comprise the majority of the so-called tip of the tongue experiences (e.g., Cross & Burke, 2004) thus are likely to be preceded by hesitation. Complex or 'heavier' items are more likely to be preceded by hesitation than simple or 'lighter' linguistic items (Clark & Wasow, 1998; see also Watanabe, et al., 2008). Code-switching between Finnish and English is also likely to create hesitation due to (a) cognitive load resulting from the speaker's language system using selective attention towards the English word while suppressing the Finnish word (e.g., Green, 1998) and (b) the speaker code-switching might increase their hesitation to subconsciously indicate to the listener that an unexpected word will be uttered (e.g., Beatie & Butterworth, 1979; Corley, et al., 2007, Tannenbaum, et al., 1965). Thus, if we find that *ee and öö* more commonly precede compound nouns, proper names and code-switching than *tota*, this

would indicate that *tota* does not have as strong a hesitation function as *ee* and *öö*.

This analysis will inform us as to whether *tota*, *ee* and *öö* are similar kinds of linguistic items.

2.1 Method

2.1.1 Corpus

We used ArkiSyn Database of Finnish Conversational Discourse (University of Turku, Department of Finnish and Finno-Ugric Languages, 2017). It contains 326,946 tokens (50,150 sentences; 27 recorded and transcribed dialogs of Finnish native speakers; 30-hours of naturalistic conversation collected from 155 adults). ArkiSyn consists of naturalistic adult spoken interaction in form of informal dialogues and group discussions.

We used the annotations made by Arkisyn automatic parser to disambiguate between the filled pause *tota* and pronoun *tota*. This corpus contained 1,306 instances of *tota* of which 106 were tagged as pronouns; 1,200 were tagged as filled pauses. Only those that were tagged as filled pauses were included in the analyses. The first author (a native Finnish speaker) then manually coded 11 out of 27 transcriptions in the Arkisyn corpus for instances of the filled pause *tota* identified by an automatic parser. The agreement between the first author and parser was strong (k = 0.86).

2.1.2 Analysis

We conducted three analyses in which we compared words that in the ArkiSyn corpus followed the filled pauses *tota* to those that followed the filled pauses *ee* and *öö*. First, we retrieved the surface frequencies of the words that followed *tota* and *ee* and *öö* in the Arkisyn corpus from Suomi24, a much bigger written corpus of Finnish (84,308,641 tokens) based on discussions of thousands of internet users (http://urn.fi/urn:nbn:fi:lb-2017021505). By the surface frequency we mean a frequency of a particular word form that followed *tota, ee* and *öö*. To give an example in English, the frequency of *going* would not include frequencies of the word forms that share the same lemma: *go, goes, went, gone*. Instead, the frequency of each surface form of the lemma (*go, goes, going, went, gone)* was calculated separately. For the surface frequency analysis we used generalised mixed-effects model (Bates, et al., 2015) in which frequency was used as a fixed effect whereas words and recordings were used as random effects. The response variable was binomial (hence generalised regression model) with the two values: either *ee and öö* (coded as 1) or *tota* (coded as 0).

Second, we investigated the parts of speech that *tota* vs. *ee* and *öö* preceded, to see if *tota* vs. *ee* and *öö* preceded different word types. We used the ArkiSyn automatic parser's annotations and coded each word that followed *tota* vs. *ee and öö* as being a noun, pronoun, verb, adjective, numeral, interjection, conjunction, adverb or end of speaker turn. We then calculated the frequency of each word type. For statistical analyses we used Pearson's Chi-squared test.

Third, we coded each noun that followed *tota* vs. *ee* and *öö* as either a 'basic' noun (e.g., *chair*), a proper noun (e.g., *Coca-Cola* or *John*), a compound noun (e.g., *child care*), or a noun produced in English instead of Finnish (and which was not a conventionally accepted loan word; e.g., if a person said *teibl* instead of *pöytä* 'table'). For this analysis we used a similar generalised mixed-effects model as for the surface frequency analysis described above.

2.2 Results

We found that the surface frequency of the succeeding words (as a fixed effect in a generalized mixed-effects model) does not predict the distinction between *ee and öö* and *tota* (p= 0.71).

Table 1 shows the distribution of the words according to the part of speech they represent. Pearson's Chi-squared test showed there was a significant difference between *tota* and *ee and öö* (p<0.001) meaning that overall the word types that *tota* and *ee and öö* precede are different. The standardized residuals suggest that in nouns, pronouns, adjectives, and interjections, the distribution is equal. However, there are more than should be expected verbs, numerals, and conjunctions that follow *ee and öö* (and less than expected verbs, numerals, and conjunctions following *tota*). On the other hand, there are more than should be expected adverbs and end of turns that follow *tota* (and less than expected adverbs and end of turns following *ee and öö*).

Table 1. Distribution of the words (N) that follow filled pauses *ee and öö* (left column) and *tota* (right column) according to the part of speech they represent (in parenthesis we show standardised residuals of the Pearson's Chi-squared test; residuals that are more than 2 mean that the number of words is significantly greater than should be expected whereas residuals less than -2 mean that the number is significantly smaller than should be expected).

POS	*ee/öö*	*tota*
noun	177 (1.94)	139 (-1.94)
pronoun	215 (-0.51)	217 (0.51)
verb	203 (5.68)	104 (-5.68)
adjective	45 (1.73)	29 (-1.73)
numeral	26 (3)	8 (-3)
interjection	236 (0.03)	227 (-0.03)
conjunction	86 (3.16)	48 (-3.16)
adverb	186 (-6.22)	301 (6.22)
end of turn	156 (-3.42)	210 (3.42)

Even though we did not find a significant difference for nouns overall (see Table 1), we coded separately those nouns that were compound nouns (N=49), proper nouns (N=81) and nouns produced in English (instead of Finnish, N=5) as we assumed these are likely to be preceded by hesitation. Table 2 shows that people were more likely to use *ee and öö* before these types of words than *tota*.

Table 2. Generalised (family = binomial) mixed-effects model with *ee and öö* vs. *tota* as a dependent variable and nouns (with four values: basic noun, compound, proper noun, and English noun) as an explanatory variable.

Fixed effects	Estimate	Std.Error	z-value	p-value
(Intercept)	-0.468	0.34	-1.375	0.169
Compound	1.211	0.414	2.926	0.003
Proper noun	1.092	0.346	3.155	0.002
English noun	2.381	1.198	1.987	0.002
Random effects				
Groups	Name	Variance	Std.Dev.	
Word	(Intercept)	<0.001	0.002	
Recording	(Intercept)	1.905	1.38	
Number of obs.: 316				
Words: 308				
Recordings: 27				

2.3. Discussion

To see if *tota* and *ee and öö* have different distributions in a naturalistic interaction in terms of the items they occur with, we investigated the frequency of the words that followed *tota* vs. *ee and öö*.

The token frequency of the items that immediately followed *tota* vs. *ee and öö* was not significantly different. The lack of effect is likely to derive from the fact that *tota* can be used as a hesitation marker. Therefore, it is not surprising it would occur to some extent with similar items (in terms of surface frequency) as *ee* and *öö*. However, our word type analysis suggests that *tota* has a weaker hesitation function than *ee* and *öö*. This is supported by the following points. First, *ee* and *öö* were more commonly followed by items from high planning or cognitive load categories, such as numerals and verbs, than *tota*. Second, even though there was no overall difference in the frequency of *tota* vs. *ee* and *öö* being followed by a noun, a closer look at different types of nouns revealed that *ee* and *öö* were used more commonly than *tota* with compound nouns (e.g., *eläinlääkäri* 'veterinarian', lit. animal's doctor), a proper noun (e.g., the word *Helsinki* or *Alexander*), or words that were pronounced in English (e.g., tafnes 'toughness'). Compound nouns, proper nouns and English nouns are likely to be relatively difficult to retrieve thus needing a high level of computational power that might create production and/or planning related hesitation. These types of items are also likely to be difficult for the interlocutor to comprehend, thus in these contexts *ee* and *öö* might have been functioning as hesitation benefitting the listener (e.g., Arnold, et al., 2004; Barr & Seyfeddinipur, 2010; Corley, et al., 2007; Watanabe, et al., 2008). *Ee* and *öö* were also found to occur before conjunctions significantly more often than *tota*. Given that clause boundaries are major planning points and thus at a typical location for hesitations (e.g., Boomer & Dittman, 1962; Clark & Fox Tree, 2002; Maclay & Osgood, 1959) *ee* and *öö* occurring in these locations more typically than *tota* suggests that *tota* does not have as strong a hesitation function. Lastly, the fact that *tota* occurred at the ends of turns, where hesitations are unnecessary, significantly more often than *ee* and *öö*, suggests that in those contexts *tota* had a pragmatic, turn-taking function (see also Etelämäki & Jaakkola, 2009).

In sum, our study 1 suggests that *tota* does not behave exactly the same way as hesitation fillers. This is in line with studies on English, where filled pauses, in particular *um*, have been argued to have pragmatic and hesitation functions and behave more like fillers-as-words than fillers-as-symptom (e.g., Clark & Fox Tree, 2002; Fox Tree, 2001; Kirjavainen, et al., 2022; Kosmala & Crible, 2022; Tottie, 2011, 2015, 2017).

Given that *tota* does not solely seem to be a hesitation marker, the question then is: what kind of a linguistic item is it? To investigate this, we conducted Study 2.

3. Study 2: Network analysis

Study 2 investigates if usage of the planning particle with politeness and hesitation functions (*tota*) correlates with the production of (a) filled pauses with hesitation functions, (b) discourse particles with repair functions, (c) grammatical politeness suffixes, (d) lexical politeness items, and (e) hesitation and (f) repair words.

The network analysis visualizes the use of words at a speaker level given other words the speaker produced. It does not mean that those words (represented by nodes) that tend to cluster together (represented by edges between the nodes) tend to be used in a similar way by the speakers. What the network analysis shows is conversational patterns of different speakers that are superimposed on each other. For example, if a person tends to be more hesitant in their speech, they tend to use some hesitancy markers. In the same vein, if a person tends to produce a more polite speech, they tend to use items that express politeness (politeness might be part of their personality or it might be required by a conversational script/situation). These quantitative relations between politeness markers and hesitancy markers are then extracted and visualized by the network analysis. Importantly, different people might use different politeness strategies – some might use more subtle politeness markings, for example, indicate politeness grammatically by using for example the conditional (-*isi*), some might use lexical items to indicate politeness (e.g., *kiitos* ‘thank you’). The co-usage of *tota* with grammatical versus lexical items can inform us as to whether *tota* is more like a grammatical or lexical element.

3.1.1 Corpus

The same corpus was used as in Study 1.

3.1.2 Analysis

In this analysis we compared the usage of *tota* (1200), that has a hesitation and politeness and uncertainty function, to the following items (with their corpus frequency in brackets):

(a) filled pauses that have a hesitation function: *öö* (541)*, ee* (800)*, hh* (5159)*, mm* (2624)

(b) discourse particles that have a repair function: *eiku ‘*no but’ (236)*, siis* ‘that is’ (1422)*, tai* ‘or’ (1199)*, nii(n)ku(in)* ‘like’ (2738)

(c) grammatical politeness markers: conditional bound morpheme -*is*(*i*) (1742), clitic -*hAn* (955), and clitic -*pA*(*s*) (31)

(d) lexical politeness markers: *kiitos* ’thank you’ (108); *ei kestä* ’you’re welcome; not at all’ (11); *anteeksi* ‘sorry’ (19)

(e) lexical planning words: *o(d)ota* ’wait’ (23), and *mietin* ‘let me think’ (24)

(f) lexical repair word: *tavallaan* ‘kind of’ (44).[3]

We analyzed statistically how the words or grammatical features are related to each other (in other words, correlated) in the network. To put simply, if a person frequently uses a certain word, e.g., the planning particle *tota*, how likely it is that

[3] As with many linguistic items, the functions assigned to words or morphemes are not always completely neat (e.g., the repair function word *siis* and the politeness clitic -*hAn* can sometimes also be used as epistemic markers).

they would frequently use some other words (e.g., *anteeksi* ‘sorry’) or grammatical suffix (e.g., the conditional morpheme; and vice versa). We visualized the results (the relations between items) using nodes (circles) representing words or grammatical features and edges (blue or red lines) that connect nodes. The network showed in Figure 1 visualizes the intensity of relations between our items of interest by the thickness and the length of the edge between the two nodes (the thicker and shorter the edge, the stronger the correlation between the two nodes). Positive associations between the variables are depicted by blue edges, and negative ones by red edges, and they are calculated using the Spearman rank correlation coefficient. The network analysis for Figure 1 was calculated in the package *bootnet* (Epskamp, Borsboom, & Fried, 2018) by using function *estimateNetwork*. The network we built is based on the Spearman correlations between the words, clitics and other bound morphemes we chose for the current study. Line width in Figure 1 is proportional to similarity (more technically: frequency of co-occurrence) between the words. To obtain a conservative (sparse) network model with only a relatively small number of edges (to reliably explain the co-variation structure in the data), we used the least absolute shrinkage and selection operator (LASSO, Tibshirani, 1996, Jankova & Van de Geer, 2018). A tuning parameter was selected by minimizing the extended Bayesian information criterion (EBIC; Chen & Chen 2008).

3.2 Results

The Network Analysis depicted in Figure 1 shows that the planning particle *tota*, filled pauses (*ee, öö, hh*) and discourse particles with a repair function (*eiku, niinku, siis*, *tai*) are clustered separately from each other. This means that if a person uses a lot of *eiku* they also use a lot of the other repair filled pauses and if a person frequently uses *ee*, they also use a lot of the other hesitation filled pauses, but that these types of filled pauses are not in any obvious way linked with *tota* or to each other. On the other hand, *tota* is closely related to grammatical politeness markers (-*isi*, -*pAs*, -*hAn*) but did not correlate with politeness words (*kiitos* ’thank you’; *ei kestä* ’you’re welcome, not at all’; *anteeksi* ‘sorry’), hesitation words (*mietin* ‘let me think’, *oota* ‘wait’) or the repair word (*tavallaan* ‘kind of’) in the network.

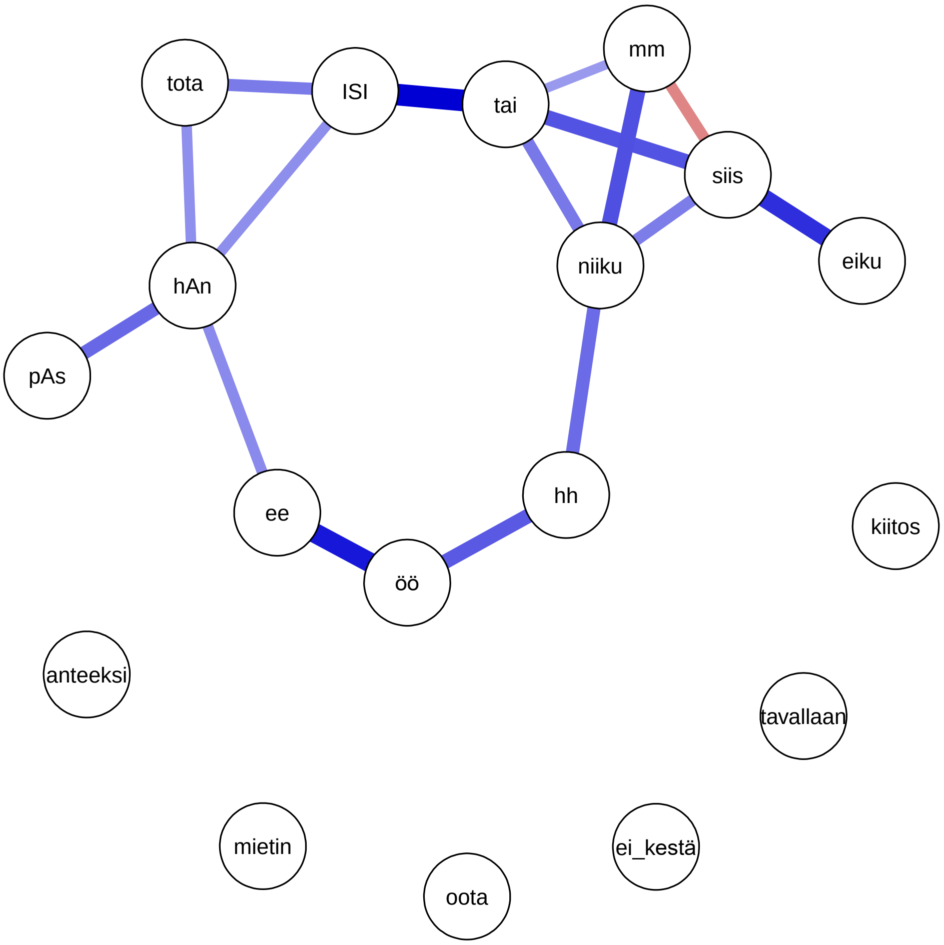

Figure 1. Estimated network of items of interest. Blue edges (lines) indicate positive relations (increase in A related to increase in B) while red edges indicate negative relations (increase in A related to decrease in B). The size and the colour intensity of edges show the intensity of the relationship.

3.3 Discussion

Study 2 found that *tota* forms a separate cluster from hesitation (*öö, ee, hh*) and repair items (*eiku, siis, niinku, tai*). This suggests that *tota* has a different function to other filled pause-like items in Finnish. This broadly fits in with previously suggested functions for Finnish filled pauses whereby items like *öö, ee and hh* have been identified as being hesitation markers, *eiku, siis, niinku* and *tai* as self-repair markers and *tota* being viewed as a planning particle (Hakulinen, et al., 2004: Table 129 §792; §861).

In the network, *tota* is closely linked with grammatical politeness markers but, interestingly, not with lexical items denoting politeness. That is, people who use a lot of grammatical politeness suffixes (e.g., the conditional suffix), also use *tota* a lot, but their use of lexical politeness items is not in any obvious way linked with the use of *tota*. One possible reason why *kiitos* ('thank you'), *anteeksi* ('sorry') and *ei kestä* ('you are welcome') do not have edges to other nodes could be their relatively low frequency of occurrence (108, 19, and 11 respectively) in our data. However, the clitic *-pA(s)* was also relatively infrequent (30), nevertheless its node still has an edge to the clitic *-hAn* (frequency 955), through which it is connected to the rest of the network. Thus, the low frequency is unlikely to explain the lack of correlation with *tota* and the lexical politeness markers.

Study 2 suggests that the three types of disfluency items (tota, hesitation, repair) are processed differently from each other, that *tota* has a politeness function and that *tota* is more similar to grammatical politeness markers than lexical politeness markers. The last claim is supported by the fact that – of all items – the only three items that are bound morphemes (grammatical suffixes *-isi*, *-hAn*, and *-pAs*) form the same cluster with the *tota*. Hence, even though technically *tota* is not a bound morpheme, it is likely to share the same grammatical function with bound morphemes *-isi*, *-hAn*, and *-pAs.*

4. General discussion

We investigated the occurrence of the planning particle *tota* in a naturalistic Finnish spoken corpus by conducting two analyses (Study 1: frequency analysis of items that immediately followed *tota*; Study 2: network analysis). Study 1 found that *tota* behaved differently from filled pauses (*ee and öö*) suggesting that, even though *tota* is a planning particle, its only function is not to mark hesitation. This is supported by fact that *tota* was found to occur more commonly in utterance final positions than filled pauses (where hesitations are unlikely). Study 2 further investigated *tota*, and tried to answer the question as to what kind of linguistic item *tota* is. Study 2 showed that *tota* grouped separately from items that have hesitation (*öö, ee, hh*) and repair (*eiku, siis, niinku, tai*) functions, supporting previous categorisations of Finnish discourse particles (Hakulinen, et al., 2004) and suggesting that disfluency markers do not form one big category but the categories are based on usage and are more nuanced. Furthermore, *tota* appeared in a cluster with grammatical politeness suffixes (*-isi, -hAn, -pAs*), but was not linked with lexical politeness items (*kiitos* 'thank you', *ei kestä* 'you are welcome', *anteeksi* 'sorry') or hesitation/repair lexemes. That is, people whose politeness strategy was to use grammatical politeness morphemes also used a lot of *tota*. This suggests that that planning markers with politeness function can be seen as being similar to grammatical items.

Our results fit in with previous studies reporting that in English filled pauses (in particular *um*) can be seen to be similar to grammatical items such as suffixes/clitics (e.g., Kirjavanen, et al., 2021; Schneider, 2014; Tottie, 2017). That is, even though filled pauses often have hesitation and repair functions (i.e., function as fillers-as-symptom filled pauses), filled pauses can also have a stronger linguistic category membership. In the case of the planning particle *tota*, this membership seems to align with grammatical items.

The idea that disfluency markers that are multifunctional such as *tota* and *um* are grammatical items is in line with the idea that while the selection of lexical items in speech production requires more awareness (e.g., declarative processing), items of grammar are often produced more automatically (procedurally) with less awareness (for a detailed explanation about the relation between procedural and non-procedural linguistic processes see e.g., Ullman 2001; 2004; 2016). Furthermore, grammatical items have relatively abstract or light meanings, that is, they add relatively little semantic content to utterances, but instead, give information for example about temporal and numerical detail, and allowing identification of participant roles. These relatively light meanings and the relatively automatic production of items like *tota* and *um* (expressing e.g., hesitation, politeness and uncertainty) can via analogy (e.g., Gentner & Medina, 1998; Tomasello, 2003) contribute to these types of item being processed similarly to grammatical items, in particular if they frequently occur in particular places in the syntactic structure or with particular words outside disfluency context (e.g., Clark & Fox Tree, 2002; Kirjavainen, et al., 2022; Schneider, 2014; Tottie, 2017).

Because the form of the planning particle *tota* is homonymous with the pronoun *tota* (*Minä haluan* ***tota*** ‘I want **that’**) it might be that it is more likely to be a filler-as-word (rather a filler-as-symptom) than a linguistic item whose form does not resemble a word (e.g., *um*, *uh*). However, the fact that we found no difference in the frequency of items that followed *tota* and *ee and öö* indicates that *tota* also has a hesitation function, regardless of its word-like form. Furthermore, as our network analysis (Study 2) showed that *tota* formed clusters with grammatical items (not lexical ones) despite its word-like form, *tota* could be comparable to less word-like hesitation markers such as *um* in English. After all, like *tota*, *um* also has hesitation and pragmatic functions (e.g. Clark & Fox Tree, 2002) and the pragmatic uses of *um* have been argued to behave like grammatical items (e.g. Kirjavainen, et al., 2022). Future investigations of the clustering effect of filled pauses with grammatical and lexical items in languages in which filled pauses have hesitation and pragmatic functions but non-word-like form would be able to shed light on this question.

Our suggestion that *tota* might be similar to grammatical items fits in well with the usage-based-constructivist viewpoint (Bybee, 1998, 2006; Bybee & Slobin 1982; Goldberg, 2006; Langacker, 2000) that assumes that all linguistic items slot into a continuum from more abstract meanings (e.g., syntax and functional items such as suffixes) to more concrete meanings (e.g., content words). According to this viewpoint, speakers build linguistic categories consisting of items that share similarities between them via processes such as analogy making (rather than linguistic items necessarily forming neat function vs. content word categories). We assume that regardless of the word-like form of *tota*, the relatively abstract politeness meaning and often the relatively automatic processing of it result in *tota* being placed towards the grammatical/functional end of the meaning continuum.

5. Conclusion

We conducted two naturalistic spoken language corpus studies that investigated the nature of Finnish filled pauses, discourse particles and the planning particle *tota*, the latter of which has hesitation and pragmatic functions. We found that the usage of *tota* was different from filled pauses and in our network analysis clustered together with grammatical politeness items (suffixes). This supports previous studies that suggest that filled pauses can, at least in some contexts, resemble and behave like grammatical items. Further study into the nature of filled pauses as grammatical items with alternative languages, items and methods is highly encouraged.

References

Aller Media ltd. 2014. The Suomi 24 Sentences Corpus (2016H2) [text corpus]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2017021505

Arnold, Jennifer, E., Michael K. Tanenhouse, Rebecca J. Altmann & Maria Fagnano. 2004. The old and thee, uh, new: Disfluency and reference resolution. *Psychological Science*, 15. 578-582.

Barr, Dale J. & Mandana Seyfeddinipur. 2010. The role of fillers in listener attributions for speaker disfluency. *Language and Cognitive Processes,* 25. 441-455.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67. 1–48.

Beattie, Geoffrey & Brian Butterworth. 1979. Contextual probability and word frequency as determinants of pauses in spontaneous speech. *Language and Speech,* 22. 201–211.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Pearson Education.

Boomer, Donald. S. & Allen T. Dittmann. 1962. Hesitation pauses and juncture pauses in speech. *Language and Speech,* 5(4). 215-220.

Bortfeld, Heather, Silvia D. Leon, Jonathan E. Bloom, Michael F. Schober & Susan E. Brennan. 2001. Disfluency rates in conversation: effects of age, relationship, topic, role and gender. *Language and Speech,* 44. 123-147.

Bosker, Hans R., Hugo Quené, Ted Sanders & Nivja H. de Jong. 2014. Native 'um's elicit prediction of low-frequency referents, but non-native 'um's do not. *Journal of Memory and Language*, 75. 104-116.

Brennan, Susan, E. & Michael, F. Schober. 2001. How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language*, 44. 274-296

Brennan, Susan. E. & Maurice Williams. 1995. The feeling of another's knowing: prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, 34. 383–398.

Chen, Jiahua & Zehua Chen. 2008. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika,* 95(3). 759–771.

Clark, Herbert H. & Jean E. Fox Tree. 2002. Using *uh* and *um* in spontaneous speaking. *Cognition,* 84. 73-111.

Clark, Herbert, H. & Thomas Wasow. 1998. Repeating Words in Spontaneous Speech. *Cognitive Psychology*, 37, 201-232.

Corley, Martin, Lucy MacGregor & David I. Donaldson. 2007. It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition,* 105. 658-668.

Corley, Martin and Oliver W. Stewart. 2008. Hesitation Disfluencies in Spontaneous Speech: The Meaning of *um*. *Language and Linguistics Compass*, 2. 589-602. https://doi.org/10.1111/j.1749-818X.2008.00068.x

Crible, Ludivine, Liesbeth Degand & Gaëtanelle Gilquin. 2017. The clustering of discourse markers and filled pauses: A corpus-based French-English study of (dis)fluency. *Languages in Contrast,* 17. 69-95.

Cross, Emily. S. & Deborah M. Burke. 2004. Do alternative names block young and older adults' retrieval of proper names? *Brain and language*, 89 (1). 174-181.

Epskamp, Sacha, Denny Borsboom & Eiko I. Fried. 2018. Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods,* 50(1). 195–212.

Etelämäki, Marja & Minna Jaakkola. 2009. 'Tota' ja puhetilanteen todellisuus. *Virittäjä*, 2/2009. 188-212.

Firth, John R. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis.* Oxford: Philological Society. 1–32.

Fox Tree, Jean. 1995. The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language,* 34. 709-738.

Fox Tree, Jean. 2001. Listeners' uses of *um* and *uh* in speech comprehension. *Memory & Cognition,* 29. 320–326.

Fox Tree, Jean E. 2002. Interpreting pauses and ums at turn exchanges. *Discourse Process*. 34 (1). 37–55.

Gentner, Dedre, & José Medina, 1998. Similarity and the development of rules. Cognition, 65(2-3), 263–297.

Green, David W. 1998. Mental control of the bilingual lexico-semantic system. *Language and Cognition*, 1, 67–81. doi:101017S1366728998000133

Götz, Sandra. 2013. *Fluency in native and non-native English speech.* Amsterdam: John Benjamins.

Haakana, Markku & Laura Visapää. 2014. Eiku – korjauksen partikkeli? *Virittäjä*, 1/2014. 41-71.

Hakulinen, Auli, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja R. Heinonen & Irja Alho. 2004. *Iso suomen kielioppi.* SKS:n toimituksia 950. Helsinki: Suomalaisen Kirjallisuuden Seura.

Harris, Zellig, S. 1954. Distributional structure. *Word*, 10(2-3). 146–162.

Irvine, Christina A., Inge-Marie Eigsti & Deborah A. Fein. 2016. Uh, um, and autism: Filler disfluencies as pragmatic markers in adolescents with optimal outcomes from autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 46(3). 1061–1070.

Janková, Jana & Sara van de Geer. 2018. Inference for high-dimensional graphical models. In Marloes Maathuis, Mathias Drton, Steffen Lauritzen & Martin Wainwright (eds.), *Handbook of graphical models*, 325–350. CRC Press: Boca Raton, Florida, USA.

Jansson-Verkasalo, Eira., Maarit Silvén, I. Lehtiö & Kurt Eggers, 2021. Speech disfluencies in typically developing Finnish-speaking children – preliminary results. *Clinical Linguistics & Phonetics,* 35, 707-726.

Kirjavainen, Minna, Ludivine Crible & Kate Beeching. 2022. Are filled pauses represented as linguistic items? Investigating the effect of exposure on the perception and production of *um*. *Language and Speech* 65. 263-289. https://doi.org/10.1177/00238309211011201

Kjellmer, Göran. 2003. Hesitation. In defence of er and erm. *English Studies*, 84, 170-198.

Kosmala Loulou & Ludivine, Crible. 2022. The dual status of filled pauses: Evidence from genre, proficiency and co-occurrence. *Language and Speech*, 65, 216-239. https://doi.org/10.1177/00238309211010862

Laakso, Minna & Marjo Lehtola. 2003. Sanojen hakeminen afaattisen henkilön ja läheisen keskustelussa. *Puhe ja Kieli*, 23, 1-24.

Levelt, Willem. J. M. 1983. Monitoring and self-repair in speech. *Cognition*, 14. 41-104.

Loy, Jia E., Hannah Rohde & Martin Corley. 2017. Effects of disfluency in online interpretation of deception. *Cognitive Science*, 41. 1434-1456.

Loy, Jia. E., Hannah Rohde & Martin Corley. 2018. Cues to Lying May be Deceptive: Speaker and Listener Behaviour in an Interactive Game of Deception. *Journal of Cognition*, 1(1). 42.

McGregor, Karla K. & Rex R. Hadden. 2020. Brief Report: "Um" fillers distinguish children with and without ASD. *Journal of Autism and Developmental Disorders*, 50(5). 1816–1821.

Maclay, Howard & Charles E. Osgood. 1959. Hesitation phenomena in spontaneous English speech. *Word*, 15. 19-44.

Menn, Lise & Loraine K. Obler. 1990. Cross-language data and theories of agrammatism. In Menn, L., & Obler, L.K. (Eds.), *Agrammatic aphasia: A cross-language narrative sourcebook*, Vol. 2 (pp. 1369–1389). Amsterdam: John Benjamins.

Mitchell, Melanie. 2019. *Artificial intelligence: A guide for thinking humans*. Penguin UK.

Newman, Mark E. J. 2010. *Networks: an introduction*. Oxford University Press, Oxford, UK.

Norrick, Neal R. 2015. Interjections. In Aijmer, K. & Rühlemann, C. (Eds.) *Corpus pragmatics : A handbook* (pp. 291-325). Cambridge : Cambridge University Press.

Penttilä, Nelly & Anna-Maija Korpijaakko-Huuhka. 2019. Disfluencies in typical Finnish-speaking adults. *The Phonetician*, 116, 28-41.

Penttilä, Nelly, Anna-Maija Korpijaakko-Huuhka & Judit Bona. 2022. Disfluency clusters in typical and atypical Finnish adult speech. A pilot study. Clinical Linguistics & Phonetics, 36, 1-16. https://doi.org/10.1080/02699206.2021.1924861

Rendle-Short, Johanna. 2004. Showing structure: using um in the academic seminar. *Pragmatics*, 14. 479-498.

Reynolds, Allan and Allan, Paivio. 1968. Cognitive and emotional determinants of speech. *Canadian Journal of Psychology,* 22 (3): 164–175.

Schegloff, Emanuel, A. 2010. Some other "uh(m)"s. *Discourse Processes,* 47. 130/174.

Schnadt, Michael J. & Martin Corley. 2006. The influence of lexical, conceptual and planning based factors on disfluency production. In *Proceedings of the twenty-eighth meeting of the cognitive science society*. Canada: Vancouver.

Schneider, Ulrike. 2014. *Frequency, hesitations and chunks. A usage-based study of chunking in English.* PhD Thesis, Albert-Ludwigs-Universität, Germany.

Shriberg, Elizabeth E. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Doctoral dissertation, University of California at Berkeley, CA.

Smith, Vicki L. & Herbert H. Clark. 1993. On the course of answering questions. *Journal of Memory and Language*, 32. 25-38.

Sorjonen, Marja-Leena & Minna Laakso. 2005. Katko vai eiku? Itsekorjauksen aloitustavat ja vuorovaikutustehtävät. *Virittäjä*, 2/2005. 244-271.

Swerts, Marc. 1998. Filled pauses as markers of discourse structure. *Journal of Pragmatics,* 30. 485–496.

Swerts, Marc & Emiel Krahmer. 2005. Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, 53(1). 81-94.

Tibshirani, Robert. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1). 267–288.

Tomasello, Michael. (2003). Constructing a language; A usage-based theory of language acquisition. Harvard University Press.

Tottie, Gunnel. 2011. Uh and um as sociolinguistic markers in British English. *International Journal of Corpus Linguistics,* 16. 173-197.

Tottie, Gunnel. 2015. Uh and um in British and American English: Are they words? Evidence from co-occurrence with pauses. In N. Dion, A. Lapierre & R. Torres Cacoullos (eds)*, Linguistic variation: Confronting Fact and Theory,* New York, Routledge: 38-54.

Tottie, Gunnel 2016. Planning what to say: Uh and um among pragmatic markers. In G. Kaltenböck, E. Keizer, & A. Lohmann (Eds.) *Outside the clause. Form and function of extra-clausal constituents* (pp. 97-122). John Benjamins.

Tottie, Gunnel. 2017. From pause to word: *uh, um* and *er* in written American English. *English Language & Linguistics,* 23. 105-130.

University of Turku, Department of Finnish and Finno-Ugric Languages. 2017. ArkiSyn Database of Finnish Conversational Discourse, Helsinki Korp Version [speech corpus]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2017022801

Ullman, Michael. 2001. The declarative/procedural model of lexicon and grammar. *Journal of Psycholinguistic Research*, 30, 37–69.

Ullman, Michael. 2004. Contributions of memory circuits to language: The declarative/ procedural model. *Cognition*, 92, 231–270.

Ullman, Michael. 2016. The declarative/procedural model: A neurobiological model of language learning, knowledge and use. In G. Hickok & S. A. Small (Eds.), *The neurobiology of language* (pp. 953–968). Elsevier.

Watanabe, Michiko, Keikichi Hirose, Yasuharu Den & Nobuaki Minematsu. 2008. Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners. *Speech Communication,* 50. 81-94.