Jonatan Moberg

SQL-NLP:N TUTKIMUSKENTTÄ: SYSTEMAATTINEN KIRJALLISUUSKARTOITUS

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2023

TIIVISTELMÄ

Moberg, Jonatan
SQL–NLP:n tutkimuskenttä: systemaattinen kirjallisuuskartoitus
Jyväskylä: Jyväskylän yliopisto, 2023, 40 s.
Tietojärjestelmätiede, kandidaatintutkielma
Ohjaaja(t): Taipalus, Toni

SQL-NLP:llä viitataan luonnollisen kielen käyttöön SQL-kielessä. SQL (engl. Structured Query Language) on ohjelmointi- ja kyselykieli, jota käytetään relaatiotietokantojen yhteydessä. NLP (engl. Natural Language Processing) tarkoittaa luonnollisen kielen prosessointia. Luonnollinen kieli viittaa ihmisten käyttämään kieleen, kuten englantiin tai suomeen. Kandidaatin tutkielma toteutettiin käyttämällä systemaattisen kirjallisuuskartoituksen menetelmää. Menetelmä valittiin, koska aiempaa tutkimusta ei ollut tehty systemaattisena kirjallisuuskartoituksena. Systemaattisen kirjallisuuskartoituksen tarkoituksena on kartoittaa aiempia tutkimuksia ja löytää mahdollisia tutkimusaukkoja. Tämän tutkielman tarkoituksena oli selvittää, kuinka paljon tutkimusta on tehty luonnollisen kielen hyödyntämisessä SQL-kielessä. Lisäksi haluttiin selvittää, millä julkaisufoorumeilla artikkeleita on julkaistu ja mistä näkökulmista artikkelit käsittelevät aihetta. Tutkielman aineisto saatiin luomalla tietokantakyselyitä neljässä eri tietokannassa. Tietokantakyselyjen tukena käytettiin taaksepäin suuntautuvaa lumipallomenetelmää. Näiden kahden menetelmän avulla tutkimusaineistoksi valikoitui 110 tieteellistä vertaisarvioidulla foorumilla julkaistua artikkelia. Tutkimuksen tulokset jaoteltiin julkaisuvuoden ja julkaisufoorumin mukaisesti. Lisäksi tutkittiin tutkimusten näkökulmaa koskien SQL-NLP:tä. Tutkielmat jaoteltiin kolmeen eri luokkaan näkökulman perusteella, joista ensimmäinen oli uudet lähestymistavat, menetelmät ja mallit, toinen kahden tai useamman menetelmän vertailu ja kolmas muut julkaisut. Tuloksista käy ilmi, että tutkimus on ajankohtaista ja uusia tutkimuksia tuotetaan koko ajan lisää. Lisäksi havaittiin, että tutkimuksia on julkaistu monipuolisesti eri julkaisufoorumeilla. Julkaisujen näkökulmista todettiin, että suuri osa julkaistuista artikkeleista esittelevät uuden luonnollisen kielen menetelmän tai tekniikan soveltamisen SQL-kielelle. Jatkotutkimuskohde aiheeseen voisi olla erilaisten SQL-NLP-sovelluksien ja -menetelmien kriittinen laajamittainen vertailu.

Asiasanat: SQL, Structured Query Language, NLP, luonnollisen kielen prosessointi, relaatiotietokannat, systemaattinen kirjallisuuskartoitus

ABSTRACT

Moberg, Jonatan
The Research Field Of SQL-NLP: A Systematic Mapping Study
Jyväskylä: University of Jyväskylä, 2023, 40 pp.
Information Systems, Bachelor's Thesis
Supervisor(s): Taipalus, Toni

The term SQL-NLP pertains to the utilization of Natural Language Processing (NLP) within the SQL (Structured Query Language) programming and query language, which is employed in the relational database domain. Natural language refers to the idiom employed by human beings, such as English or Finnish. This bachelor's thesis was conducted as a systematic mapping study methodology. This methodology was elected due to the absence of prior research conducted as a systematic mapping study. The prime objective of a systematic mapping study is to map antecedent research and unearth prospective research gaps. The objective of the thesis was to ascertain the extent of extant research with regards to the utilization of natural language in SQL, as well as to determine the publication fora and perspectives from which relevant articles approach the topic. The material for this thesis was garnered through the creation of database queries in four distinct databases. The backward snowballing method was utilized to complement these queries. The techniques culminated in the selection of a research sample consisting of 110 scientific articles published in peer-reviewed fora. The results of the study were classified according to year of publication and publication forum. Furthermore, the study probed the SQL-NLP perspective of the articles. The articles were divided into three distinct categories, predicated on their respective perspectives. The first category pertained to novel approaches, methodologies, and models, the second category focused on the comparison of two or more methodologies, and the third category encapsulated publications of an alternative nature. The findings indicate that research on the subject is germane, and that novel research is being produced on a constant basis. Furthermore, it was discovered that research had been published across a range of publication fora. Regarding the perspectives of the articles, a significant proportion of the articles proffered novel natural language methodologies or techniques, which were applied to SQL. A potential future research avenue on this subject could be a large-scale, critical comparison of divergent applications and methodologies of SQL-NLP.

Keywords: SQL, Structured Query Language, NLP, Natural Language Processing, relational databases, a systematic mapping study

KUVIOT

TAULUKOT

SISÄLLYS

1 JOHDANTO

Tässä kandidaatintutkielmassa tutkitaan luonnollisen kielen prosessoinnin eli NLP:n (Natural Language Processing) hyödyntämistä SQL-kielessä (engl. Structured Query Language) systemaattisen kirjallisuuskartoituksen avulla. Tutkimuksen tarkoituksena on kartoittaa aiempaa tutkimusta aiheesta ja saada tietoa tutkimusten lähestymistavoista sekä mahdollisista tutkimusaukoista. Lisäksi tutkielman tavoitteena on saada laaja kuva tutkimuksen määrästä ja näkökulmasta koskien SQL–NLP:tä. Tutkimuksessa keskitytään viimeisen viiden vuoden aikana julkaistuihin tutkimuksiin, sillä SQL–NLP ja tekoäly ovat ottaneet suuria harppauksia viime vuosien aikana.

Tutkimus tehdään systemaattisena kirjallisuuskartoituksena pääosin Taipaluksen (2023) sekä Petersenin, Vakkalankan ja Kuzniarzin (2015) ohjeistusta noudattaen. Tutkimusmenetelmäksi valikoitu systemaattinen kirjallisuuskartoitus, koska kyseistä menetelmää ei olla aiemmin käytetty SQL–NLP:stä tehtyyn tutkimukseen. Tutkielman tutkimuskysymykset ovat:

1. Kuinka paljon tutkimusta SQL-NLP:stä on tehty viimeisen viiden vuoden aikana?
2. Millä foorumeilla tutkimusta SQL-NLP:stä on julkaistu?
3. Mistä näkökulmista SQL-NLP:tä käsitellään?

SQL-NLP:llä tarkoitetaan SQL-kyselykielen, käyttämistä ja toteuttamista luonnollisen kielen prosessoinnin avulla. SQL on kysely- ja ohjelmointikieli, jota käytetään relaatiotietokantojen hallintaan ja tietojen käsittelyyn. Luonnollisen kielen avulla pyritään helpottamaan niiden käyttöä.

Pystyäkseen hakemaan tietoa tietokannasta, käyttäjän on osattava SQL-kielen syntaksi (Uma, Sneha, V., Sneha, G., Bhuvana & Bharathi, 2019). Kuten Uma ym. (2019) esittävät, luonnollisen kielen prosessointi tähtää siihen, että tieto olisi helpommin jokaisen saatavilla riippumatta siitä, onko riittävää osaamista muodostaa kyselykielen mukainen tietokantakysely. Lisäksi käyttäjän tulee ymmärtää tietokannan taulujen rakenne ja niiden väliset yhteydet, jotta tietokantakyselyiden luominen on mahdollista. Tämä voi tuottaa vaikeuksia henkilöille,

joiden osaaminen tietotekniikassa ja SQL-kielessä ei ole riittävällä tasolla (Karimi, Rasel & Abdullah, 2022). Etenkin englantia osaamattomille SQL-kielen ymmärtäminen ja opetteleminen aiheuttaa haasteita, sillä SQL-kieli pohjautuu englantiin.

NLP eli luonnollisen kielen prosessointi tarkoittaa, että esimerkiksi puhuttua tai kirjoitettua kieltä pystyttäisiin käsittelemään. Tietokantojen ja SQL-kyselykielen kontekstissa tämä tarkoittaa esimerkiksi sitä, että osaamista SQL-kielen syntaksista ei tarvittaisi, vaan ihmisille luontevammalla kielellä pystyttäisiin suorittamaan samat toiminnot kuin kyselykielellä (Karimi ym., 2022).

Lähes jokaisella toimialalla relaatiotietokannat ovat keskiössä datan tallentamisessa ja hakemisessa (Deshpande, Kothari, Salvi, Mane & Kolhe, 2022). Lisäksi suuri osa informaatiosta on tietokantojen muodossa (Uma ym., 2019). Tämän myötä SQL-kielen osaaminen ja kyky kommunikoida tietokantojen kanssa on tärkeämpää kuin koskaan.

Tutkielman aineisto saadaan luomalla tietokantahakuja ja käyttämällä taaksepäin suuntautuvaa lumipallomenetelmää (engl. backward snowballing). Tämän jälkeen saatuja artikkeleita verrataan keskenään. Aineistoa luokitellaan tutkimuskysymysten mukaisesti julkaisufoorumin, julkaisuvuoden ja tutkimusnäkökulman osalta. Luokittelun perusteella pystytään luomaan johtopäätöksiä mahdollisista tutkimusaukoista ja tarvittavista uusista tutkimuksista.

Tutkimuksen tuloksista ilmenee, että aihe on ajankohtainen myös julkaistujen artikkelien määrän perusteella. Tutkimukseen valikoituneista artikkeleista suuri osa keskittyy uuden SQL-NLP -mallin tai -lähestymistavan esittelemiseen. Osassa artikkeleista verrataan aiempia kehiteltyjä menetelmiä keskenään, mutta vertaileville tutkimuksille olisi tilaa tutkimuskentällä enemmänkin.

Tutkielman toisessa luvussa perehdytään SQL-kielen ja luonnollisen kielen prosessoinnin ominaisuuksiin. Kolmannessa luvussa esitetään systemaattinen kirjallisuuskartoitus tutkimusmenetelmänä ja kuvataan tutkielman toteutus vaiheittain. Neljännessä luvussa esitetään saatuja tuloksia ja luokitellaan valikoituneet artikkelit tutkimuskysymysten mukaisella tavalla. Viidennessä luvussa pohditaan ja analysoidaan saatuja tuloksia sekä tehdään niiden perusteella johtopäätöksiä. Tutkielman kuudes ja viimeinen luku on yhteenveto, jossa käsitellään mahdollisia jatkotutkimusaiheita.

2 SQL-NLP

SQL-NLP on yhdistelmä SQL-kielen ja NLP:n ominaisuuksia. SQL on yleisesti käytetty tietokantakyselyjen kieli, joka mahdollistaa tietojen hakemisen, päivittämisen ja hallinnan tietokannoissa. NLP puolestaan tarkoittaa luonnollisen kielen käsittelyä, jonka avulla pyritään mahdollistamaan tietokoneiden ymmärtäminen ja käsittely ihmisten käyttämällä kielellä, kuten suomen tai englannin kielellä (Shah, Das, Shahane, Parikh & Bari, 2021).

SQL-NLP voi tarkoittaa erilaisia asioita riippuen siitä, miten NLP-ominaisuuksia käytetään SQL-kyselyissä. Yksi mahdollinen sovellutus on luonnollisen kielen kyselyjen salliminen tietokantoihin. Tällöin käyttäjä voi esittää kysymyksen tietokannasta luonnollisella kielellä ja SQL-NLP -järjestelmä muuntaa sen automaattisesti SQL-kyselyksi, joka suoritetaan tietokannassa. Tämän myötä käyttäjän ei tarvitse osata SQL-kielen syntaksia, vaan pystyy kommunikoimaan tietokantojen kanssa ilman suurta työmäärää (Uma ym., 2019).

Toinen mahdollinen sovellutus on NLP:n käyttäminen tietokantakyselyjen luomisessa esimerkiksi puheen avulla. Tällöin tietokantojen ja luonnollisen kielen tekniikkaa yhdistetään puheentunnistusteknologiaan (Shah, Li, Kumar & Saul, 2020). SQL-NLP-järjestelmä voi ymmärtää puhutun kysymyksen luonnollista kieltä ja pyrkiä optimoimaan sen SQL-kyselyksi, joka suoritetaan nopeasti ja tehokkaasti tietokannassa.

SQL-NLP:tä pystytään myös soveltamaan muihin eri tietojärjestelmätieteenalan osa-alueisiin ja tehtäviin, kuten tietokantadatan luokitteluun, datan analysointiin ja SQL-kielen oppimiseen. Uusien teknologioiden myötä luonnollisen kielen hyödyntäminen tietokannoissa nopeuttaa ja helpottaa tietokantojen käyttöä. Kim, So, Han ja Lee (2020) toteavat tutkimuksessaan, että viime vuosina luonnollisen kielen kääntämistä SQL-kieleksi on aktiivisesti kehitetty niin tietokanta- kuin luonnollisen kielen yhteisön puolelta. Käyttöliittymää, jota käytetään kääntämään luonnollista kieltä SQL-kieleksi, kutsutaan nimellä tietokannan käyttöliittymä luonnolliselle kielelle (engl. Natural Language Interface to Database, NLIDB) (Reshma & Remya, 2017).

Yksi keskeisimmistä haasteista luonnollisen kielen kääntämisessä SQL-kieleen on sanojen monimerkityksisyys (Katsogiannis-Meimarakis & Koutrika,

2021). Esimerkiksi suomen kielen sana *kurkku* voi tarkoittaa joko vihannesta tai ruumiinosaa. Toiseksi merkittäväksi ongelmaksi Katsogiannis-Meimarakis ja Koutrika (2021) luettelevat sanojen erot luonnollisessa kielessä ja tietokannassa. Tällä tarkoitetaan sitä, että luonnollisen kielen sana ei vastaa tietokannassa olevaa sanaa, vaikka sanalla olisikin sama merkitys. Näihin haasteisiin pyritään vastamaan tulevaisuudessa uusilla SQL-NLP teknologioilla.

2.1 SQL-kieli

SQL (engl. Structured Query Language) on laajimmin käytetty kieli datan hallitsemiseen tietokannoissa (Silva, Almeida & Michell, 2016). SQL-kieltä käytetään erityisesti relaatiotietokannoissa, joissa tietoa tallennetaan taulujen muodossa. SQL on standardoitu kieli, jonka ANSI (American National Standards Institute) ja ISO (International Organization for Standardization) ovat kehittäneet (Silva ym., 2016). SQL kielen peruskomentoja ovat SELECT, INSERT, UPDATE ja DELETE, jotka mahdollistavat tietojen hakemisen, lisäämisen, päivittämisen ja poistamisen tietokannoista. SQL kielen avulla voidaan myös luoda tauluja, indeksejä ja muita tietokannan rakenteita.

Ensin luodaan tietokanta, joka sisältää yhden tai useamman taulun. Taulut koostuvat sarakkeista (attribuuteista) ja riveistä (tietueista). Kun tietokanta on luotu, tietueita voidaan lisätä tauluun INSERT-komennolla. Komennossa määritellään tietueen arvot ja sarakkeet, joihin ne kuuluvat. Tietueita voidaan poistaa DELETE-komennolla. Tässä määritellään tietueiden ehdon mukaan, mitkä tietueet poistetaan. Tietueita voidaan päivittää UPDATE-komennolla. Tässä määritellään tietueiden ehdon mukaan, mitkä tietueet päivitetään ja mitkä arvot annetaan. Tietoja voidaan hakea SELECT-komennolla. Tämä komento mahdollistaa tietojen hakemisen tietokannasta tietyin ehdoin ja järjestyskriteerein.

SQL-komentoja käytetään yleisesti sovelluskehittäjien ja tietokanta-analyytikoiden keskuudessa tietokantojen käytön ja hallinnan yksinkertaistamiseksi. Se on erittäin tärkeä työkalu suurien tietokantojen tehokkaassa käytössä ja sen käyttö on laajalle levinnyt kaikkialle tietokantojen ylläpitämisessä ja hallinnassa. Tietokantojen käyttö on yleistynyt myös muiden käyttäjien osalta, koska tietokannat sisältävät suuren osan tämän päivän informaatiosta (Uma ym., 2019).

SQL-kieltä käytetään laajalti erilaisissa sovelluksissa ja järjestelmissä, joissa käsitellään suuria määriä tietoa, kuten esimerkiksi pankki- ja talousjärjestelmissä, verkkosivustoilla, verkkokaupoissa, asiakastietojärjestelmissä ja terveydenhuollon sovelluksissa. Myös tietojen analysoinnissa ja raportoinnissa SQL on tärkeä työkalu, sillä sen avulla voidaan tehdä monipuolisia kyselyjä tietokannasta ja käsitellä suuria datamääriä tehokkaasti.

2.2 NLP

Kłosowskin (2018) määrittelyn mukaan NLP (Natural Language Processing) eli luonnollisen kielen prosessointi tai luonnollisen kielen käsittely tarkoittaa tietokoneiden kykyä käsitellä ja ymmärtää ihmisten käyttämää kieltä, kuten puhetta ja kirjoitettua tekstiä. Wilson, Martin ja Gilbert (2010) määrittelevät, että luonnollisen kielen prosessointi kuuluu osaksi tekoälyn ja kielitieteen tutkimuskenttiä. Toisaalta Shah, Das, Shahane, Parikh ja Bari (2021) määrittelevät luonnollisen kielen prosessoinnin olevan tekoälyn tutkimushaara. He lisäävät, että NLP:n pohjimmainen tarkoitus on mahdollistaa kommunikaatio ihmisten ja tietokoneiden välillä ilman monimutkaisia ja vaikeita toimenpiteitä (V. Shah ym., 2020).

Luonnollisen kielen prosessointia hyödynnetään tietojen hakemiseen, konekääntämiseen ja kielen analyysiin (V. Shah ym., 2020). Toisaalta Sanyal, Shukla ja Agrawal (2021) toteavat, että luonnollista kieltä prosessoivaa teknologiaa käytetään kaikilla aloilla ihmisen ja tietokoneen välisessä vuorovaikutuksessa. He lisäävät, että luonnollisen kielen prosessointia käytetään laajasti myös muissa kuin tekstipohjaisissa sovelluksissa. Näistä tutkimuksessa annetaan esimerkkeinä ääni-, video- ja animaatiojärjestelmät (Sanyal ym., 2021).

Karimi, Rasel ja Abdullah (2022) esittävät tutkimuksessaan luonnollisen kielen tarjoamia mahdollisuuksia niille henkilöille, joilla on erityistarpeita. Esimerkiksi fyysisiä rajoitteita omaava ihminen voi tarvita luonnollisen kielen käsittelyyn erilaisia menetelmiä. Tutkielmassa luetellaan muun muassa ääniohjaus, katseenseuranta ja graafinen esitys. Sen lisäksi edellä mainittuja tekniikoita voidaan yhdistää esimerkiksi oppimisvaikeuksista kokeville henkilöille (Karimi ym., 2022).

Luonnollisen kielen prosessointia voidaan myös käyttää tiedon poiminnassa, joka tarkoittaa tietokonejärjestelmän kykyä tunnistaa tiettyjä tietoja tekstistä. Esimerkiksi, jos tietokannassa on satoja tuhansia asiakasarvioita, luonnollisen kielen prosessoinnin avulla voidaan tunnistaa tietyn tuotteen tai palvelun yleisimmät valitukset tai kehut.

Kaiken kaikkiaan NLP on tärkeä työkalu tietokannoissa, sillä se auttaa järjestelmiä käsittelemään valtavia määriä ihmisten käyttämiä kieliä ja löytämään tärkeät tiedot nopeasti ja tarkasti. Sen avulla pystytään parantamaan ihmisten ja tietokoneiden välistä vuorovaikutusta ja kommunikaatiota. Lisäksi sen kehitys on parantunut merkittävästi viime vuosikymmeninä ja luonnollisesta kielestä on noussut tärkeä tutkimusaihe tietojärjestelmätieteen alalla.

3 TUTKIMUSMENETELMÄ

Tässä luvussa kuvataan tutkielman tutkimusmenetelmä, systemaattinen kirjallisuuskartoitus, pääosin hyödyntäen Petersenin, Vakkalankan ja Kuzniarzin (2015) sekä Taipaluksen (2023) muodostamia ohjeistuksia systemaattisesta kirjallisuuskartoituksesta. Tutkimuksessa painottuu Taipaluksen (2023) ohjeistus, sillä se on uudempi ja siinä on käytetty uudempaa lähdemateriaalia, kuin Petersenin ym. (2015) julkaisussa. Seuraavassa alaluvussa verrataan systemaattista kirjallisuuskartoitusta sitä tavanomaisempaan systemaattiseen kirjallisuuskatsaukseen.

3.1 Systemaattinen kirjallisuuskartoitus

Systemaattisessa kirjallisuuskartoituksessa on paljon samoja piirteitä kuin systemaattisessa kirjallisuuskatsauksessa. Yksi keskeisimmistä eroista on tutkimuksen tavoite (Petersen ym., 2015). Kun systemaattisessa kirjallisuuskatsauksessa tähdätään syvällisempään ymmärrykseen aiheesta aiempiin tutkimuksiin perustuen, niin systemaattisessa kirjallisuuskartoituksessa tavoitellaan korkeamman tason ymmärrystä aiheeseen liittyvistä aiheista ja lähestymistavoista (Taipalus, 2023). Budgen, Turner, Brereton ja Brereton (2008) lisäävät, että systemaattisen kirjallisuuskartoituksen tavoitteena on selvittää tutkimusalueen laajuutta ja rakennetta. Petersenin ym. (2015) mukaan systemaattisen kirjallisuuskartoituksen tavoitteena on antaa yleiskuvaus aiemmista tutkimuksista ja luokitella aiheeseen liittyvää primääritutkimusta. Taipalus (2023) esittää, että systemaattisessa kirjallisuuskartoituksessa aineiston määrä on yleensä laajempi kuin systemaattisessa kirjallisuuskatsauksessa, mutta aineiston tulkinta ei ole niin laajaa. Tämä johtuu tutkimusmenetelmien tavoitteiden eroavaisuuksista.

Toinen keskeisimmistä eroista systemaattisen kirjallisuuskartoituksen ja systemaattisen kirjallisuuskatsauksen välillä Kitchenhamin, Budgenin ja Breretonin (2010) mukaan ovat tutkimuskysymykset. Systemaattisessa kirjallisuuskartoituksessa tutkimuskysymykset ovat geneerisempiä ja ne liittyvät

tutkimustrendeihin, kun taas systemaattisessa kirjallisuuskatsauksessa tutkimuskysymykset ovat paljon spesifimpiä (Kitchenham ym., 2010).

Koska tutkimuskysymykset ovat menetelmissä erilaiset, myös tulosten ilmoittaminen edellä mainituissa tutkimusmenetelmissä eroavat toisistaan. Kuten Kitchenham ja muut (2010) kertovat, systemaattisessa kirjallisuuskatsauksessa tulosten ilmoittaminen liittyy spesifiin tutkimuskysymykseen vastaamiseen. Taipalus (2023) toteaa, että systemaattisen kirjallisuuskartoituksen tulosten raportointi tapahtuu usein kahdessa eri vaiheessa. Ensimmäisessä vaiheessa raportoidaan systemaattisen prosessin vaiheet, kuten tutkimukseen valikoituvien primääritutkimuksien valintakriteerit, valitut tietokannat ja hakutermit sekä valintakriteereiden läpäisseiden artikkeleiden määrä. Toisessa vaiheessa raportoidaan systemaattisen kartoituksen tulokset, eli esimerkiksi julkaisuvuodet ja julkaisufoorumit (Taipalus, 2023).

3.2 Tutkimuskysymykset

Tutkielman tutkimuskysymykset ovat systemaattiselle kirjallisuuskartoitukselle tyypillisiä. Seuraavien tutkimuskysymysten avulla saadaan laaja käsitys SQL-NLP:n tutkimustrendeistä ja mahdollisista tutkimusaukoista.

1. Kuinka paljon tutkimusta SQL-NLP:stä on tehty viimeisen viiden vuoden aikana?
2. Millä foorumeilla tutkimusta SQL-NLP:stä on julkaistu?
3. Mistä näkökulmista SQL-NLP:tä käsitellään?

3.3 Valintakriteerit tutkimuksille

Taipalus (2023) esittää, että tutkimuksen valintakriteerit tulee esittää mahdollisimman yksiselitteisesti. Valintakriteerit voivat olla joko tutkimukseen sisällyttäviä kriteereitä tai tutkimuksesta poissulkevia kriteereitä. Negaation kautta sama kriteeri voi joko olla sisällyttävä tai poissulkeva. Taipalus (2023) lisää, että tutkimuksessa on myös mahdollista käyttää pelkästään sisällyttäviä tai poissulkevia kriteereitä. Tässä tutkimuksessa valintakriteerit tehdään tutkimukseen mukaan otettavien kriteereiden pohjalta. Kriteerit ovat seuraavalla sivulla esitetyssä taulukossa (taulukko 1).

TAULUKKO 1 Tutkimusaineiston valintakriteerit

Tunnus	Selitys
K1	Koko teksti on saatavilla ilmaiseksi Jyväskylän yliopisto-opiskelijalle
K2	Tutkimus käsittelee SQL-NLP:tä
K3	Artikkeli on julkaistu englanniksi
K4	Artikkeli on julkaistu vuosina 2018-2022
K5	Artikkeli on julkaistu vertaisarvioidulla foorumilla

3.4 Tutkimusten haku

Tässä tutkimuksessa käytettiin neljää digitaalista tietokantaa tutkimusten hakemiseen. Kyseiset tietokannat olivat AIS eLibrary, ACM Digital Library, IEEE Xplore ja ScienceDirect. Taipaluksen (2023) mukaan tyypillinen tietokantojen määrä tutkimuksessa on kahdesta viiteen. Tämän myötä tähän tutkimukseen valikoitu neljä tietokantaa. Tutkimukseen valikoituivat juuri nämä tietokannat saatavuuden vuoksi Jyväskylän yliopiston opiskelijana. Lisäksi Taipalus (2023) systemaattisen kirjallisuuskartoituksen ohjeistuksessaan mainitsee edellä mainitut tietokannat merkittävinä tietojärjestelmätieteenalan tietokantoina. Tutkimuksessa ei käytetty tietokantahakujen apuna muun muassa Google Scholarin tietokantaa, sillä se ei Taipaluksen (2023) mukaan sovellu systemaattiselle kirjallisuuskartoitukselle. Syy tälle on liian sattumanvarainen ja suuri määrä hakutuloksia. Tämän myötä tutkimus ei olisi muille tutkijoille toistettavissa (Taipalus, 2023).

Toisena menetelmänä tutkimusten hakemiseen käytetään lumipallomenetelmää. Wohlin, Kalinowski, Romero Felizardo ja Mendes (2022) määrittelevät lumipallomenetelmän (engl. snowballing) tarkoittavan uusien lähteiden etsimistä aiempia lähteitä ja sitaatteja apuna käyttäen. Petersen ym. (2015) toteavat, että yhdistämällä tietokantahaut lumipallomenetelmän kanssa, saadaan tarkemmat tulokset tutkittavasta aiheesta.

Taipalus (2023) kertoo lumipallomenetelmän jakautuvan taaksepäin suuntautuvaan (engl. backward snowballing) ja eteenpäin suuntautuvaan (engl. forward snowballing) lumipallomenetelmään. Tässä tutkimuksessa tietokantahakujen tukena käytettiin taaksepäin suuntautuvaa lumipallomenetelmää. Taaksepäin suuntautuvassa lumipallomenetelmässä jo valituista primäärituttkimuksista käydään läpi käytetyt lähteet ja niistä olennaiset sisällytetään tutkimukseen (Taipalus, 2023). Lumipallomenetelmän myötä löydetyt tutkimukset käydään läpi ja hakukriteerit läpäisevät tutkimukset sisällytetään tähän tutkimukseen. Tämän avulla varmistetaan se, että kaikki relevantit julkaisut aiheeseen liittyen saadaan sisällytettyä tutkimukseen, jotka eivät tietokantahakujen myötä välttämättä löytyneet (Wohlin ym., 2022).

3.4.1 Tietokantahaut

Ensimmäisessä tietokantahaussa valituista neljästä tietokannasta etsittiin tutkimuksia. Hakulausekkeena oli ”sql AND nlp”. Tarkempana hakukriteerinä oli rajattu julkaisuvuosi vuosien 2018-2022 välille. IEEE Xplore -tietokannassa hakulauseke rajattiin koskemaan metatekstiä eli otsikkoa, tiivistelmää tai avainsanoja. Tälle syynä oli se, että ilman rajausta kyseinen tietokanta palautti hakutuloksia jo 937. Kolmessa muussa tietokannassa termit SQL ja NLP saivat esiintyä missä kohtaa tahansa tekstiä, eikä sitä ollut rajattu koskemaan vain metatekstiä. Hakulauseketta ei ollut tutkimuksen laajuuden ja työmäärän kannalta järkevää laventaa, sillä pelkästään yksinkertaisella hakulausekkeella saatiin tuloksia 829 otsikkotason tarkasteluun. Alla olevasta taulukosta (taulukko 2) ilmenee kunkin tietokannan hakutulosten lukumäärä määrätyin kriteerein.

TAULUKKO 2 Tietokantahaun tulokset

Tietokanta	**Hakulauseke**	**Hakukriteerit**	**Hakutulosten lukumäärä**
ACM Digital Library	”sql AND nlp”	Hakusanat voivat esiintyä missä kohtaa tahansa tekstiä, julkaisuvuodet 2018-2022	462
AIS eLibrary	”sql AND nlp”	Hakusanat voivat esiintyä missä kohtaa tahansa tekstiä, julkaisuvuodet 2018-2022. Vain vertaisarvioidut artikkelit	2
IEEE Xplore	”sql AND nlp”	Hakusanat esiintyvät otsikossa, tiivistelmässä tai avainsanoissa, julkaisuvuodet 2018-2022	25
ScienceDirect	”sql AND nlp”	Hakusanat voivat esiintyä missä kohtaa tahansa tekstiä, julkaisuvuodet 2018-2022	340

3.4.2 Lumipallomenetelmä

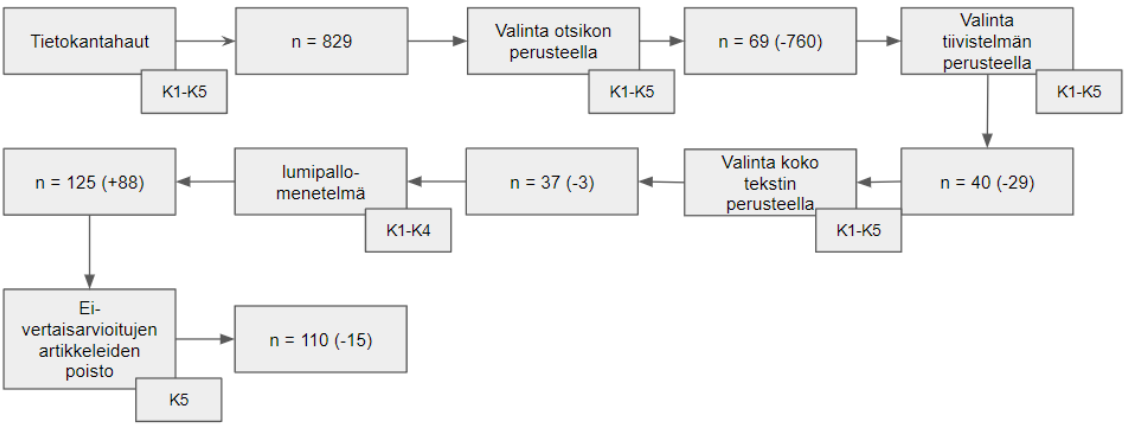
Tutkielmassa käytettiin tietokantahakujen tukena taaksepäin suuntautuvaa lumipallomenetelmää. Tämä oli merkittävä osa tämän tutkimuksen aineiston hakua, koska käytetty hakulauseke ei kattanut kaikkea luonnollisen kielen

hyödyntämistä SQL-kielen kanssa. Esimerkiksi hakulauseke ”sql AND natural language” olisi tuottanut liian suuren määrän hakutuloksia yhden tutkijan analysoitavaksi, jonka takia tietokantahakuja ei suoritettu muilla hakulausekkeilla.

Taaksepäin suuntautuvassa lumipallomenetelmässä lähdeluettelot käytiin otsikkotasolla läpi lopullisista tutkimukseen valikoituneista primääritutkimuksista. Jos otsikkotasolla ei pystytty olemaan varmoja artikkelin sopivuudesta tähän tutkimukseen, käytiin artikkelin tiivistelmä ja koko teksti tarvittaessa läpi. Tämän avulla pystyttiin varmistamaan se, että tutkijan subjektiivinen hakulausekkeiden valinta ei vääristänyt lopullista aineistoa tutkimuksessa merkittävästi. Tietokantahakujen ja lumipallomenetelmän avulla toteutettu aineiston valintaprosessi on esitettynä seuraavassa luvussa.

3.5 Tutkimusten valinta

Tietokantahakujen 829 hakutulosta käytiin ensiksi läpi otsikkotasolla. Otsikon perusteella sopivia julkaisuja löytyi 69 kappaletta. Nämä julkaisut käytiin vielä läpi tiivistelmän ja koko tekstin perusteella. Yhteensä tietokantahakujen tuloksena tutkimukseen valikoitui 37 artikkelia. Näihin 37 artikkeliin sovellettiin taaksepäin suuntautuva lumipallomenetelmää ja saaduista tuloksista poistettiin duplikaatit. Tämän jälkeen tutkimukseen oli valikoitunut yhteensä 125 julkaisua, joista poistettiin vielä ne tuotokset, joiden vertaisarvioinnista ei voitu olla varmoja. Lopulta tutkimukseen päätyi analysoitavaksi 110 julkaisua (ks. liite 1). Koko valintaprosessi on esitelty alla olevassa kuviossa (kuvio 1).



KUVIO 1 Aineiston valintaprosessi: n viittaa valittujen julkaisujen määrään. K1-K5 viittaavat tutkimusaineiston valintakriteereihin. Jokaisessa työvaiheessa on poistettu duplikaatit

3.6 Validiteettiin liittyvät uhkat

Petersenin ja muiden (2015) sekä Taipaluksen (2023) ohjeistuksien mukaisesti systemaattisen kirjallisuuskartoituksen tuottamiseen liittyy uhkia validiteetin näkökulmasta. Yksi keskeisimmistä uhkista validiteetin kannalta on aineiston valitseminen ainoastaan yhden tutkijan toimesta (Petersen ym., 2015). Tämä uhka on keskeisin validiteettiin vaikuttava tekijä tässä tutkielmassa.

Koska aihealue on tutkielmassa suhteellisen laaja ja aineistoa kertyi paljon, aineistoa tarkasteltiin useampaan kertaan kriittisesti ja iteratiivisesti. Lisäksi vääristymää pyrittiin vähentämään tarkoin valintakriteerein ja ilmoittamalla aineiston valintaprosessi mahdollisimman eksplisiittisesti. Tutkielman ohjaajaa konsultointiin muun muassa aineistohaun ja julkaisufoorumeiden osalta, jotta aineiston hankinta ja julkaisukanavat olisivat mahdollisimman selkeästi esitettynä.

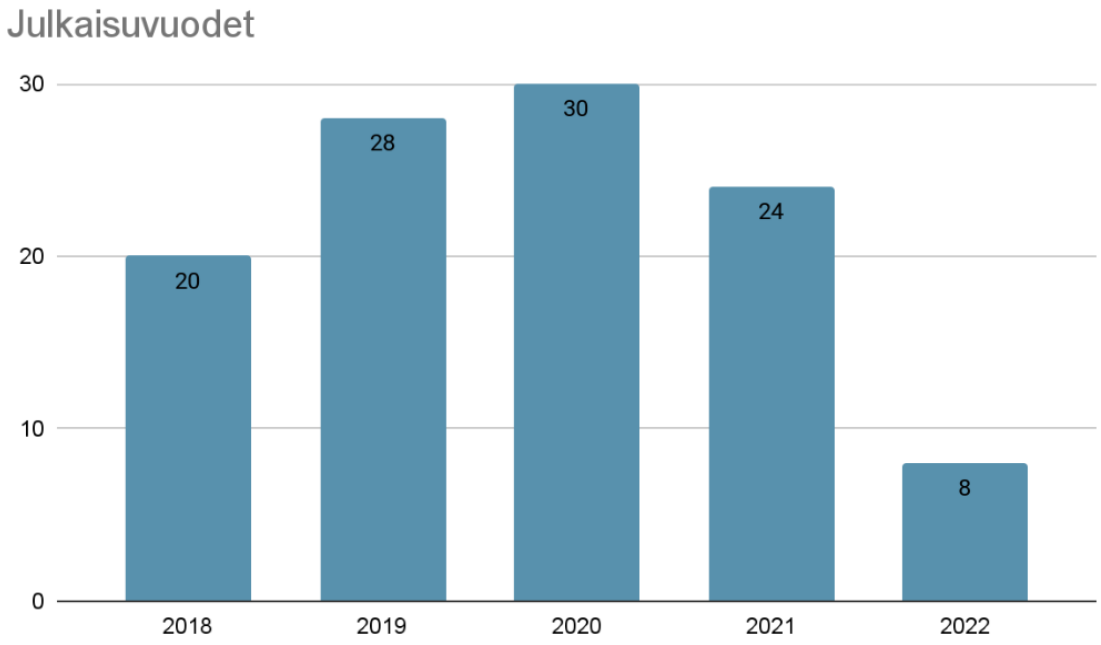
Tutkielmassa vääristymää voi aiheuttaa kirjoittajan valitsemat tietokannat ja hakulausekkeet. Tietokannoiksi valittiin Taipaluksen (2023) ohjeistuksen mukaisesti tietojärjestelmätieteen merkittävimmät tietokannat. Hakulauseke oli yksinkertainen ”sql AND nlp”, joten taaksepäin suuntautuva lumipallomenetelmä oli tarpeellinen. Eteenpäin suuntautuvaa lumipallomenetelmää harkittiin tutkielman aineistohaun menetelmänä, mutta se suljettiin pois liian suuren työmäärän vuoksi. Eteenpäin suuntautuvassa lumipallomenetelmässä (engl. forward snowballing) aineistoa etsitään uusista artikkeleista, jotka viittaavat jo valittuihin tutkimuksiin (Taipalus, 2023; Wohlin ym., 2022). Koska tätä menetelmää ei käytetty tutkielmassa, on mahdollista, että osa vuoden 2021 ja 2022 julkaisuista jäi tutkimuksen ulkopuolelle.

4 TULOKSET

Tässä luvussa luokitellaan saatu aineisto esitettyjen tutkimuskysymysten pohjalta. Aineistoksi valikoitu 110 artikkelia koskien SQL-NLP:n tutkimusta. Ensimmäiseksi esitetään julkaisuvuodet ja julkaisufoorumit. Tämän jälkeen luokitellaan julkaisut niiden lähestymistavan pohjalta.

4.1 Julkaisuvuodet

Kuviossa (kuvio 2) on esitelty aineistohausta saatujen julkaisujen julkaisuvuodet. Tutkimusaineiston julkaisuvuodet rajattiin vuosien 2018-2022 välille. Kuten luvussa Validiteettiin liittyvät uhkat – mainittiin, on mahdollista, että vuosien 2021 ja 2022 julkaisuista puuttuu joitakin tuloksia eteenpäin suuntautuvan lumipallomenetelmän poissulkemisen myötä.



KUVIO 2 Julkaisuvuodet

4.2 Julkaisufoorumit

Alla olevassa taulukossa (taulukko 3) on esitelty 110 artikkelin julkaisufoorumit. International Joint Conference on Natural Language Processing (IJCNLP) kuuluu yhdistettyyn konferenssiin, joka on järjestetty jonkin toisen konferenssin kanssa yhteisesti. Nämä kirjattiin erikseen taulukkoon, jonka myötä taulukossa on kahdeksan tulosta enemmän, kuin valikoituneita artikkeleita yhteensä.

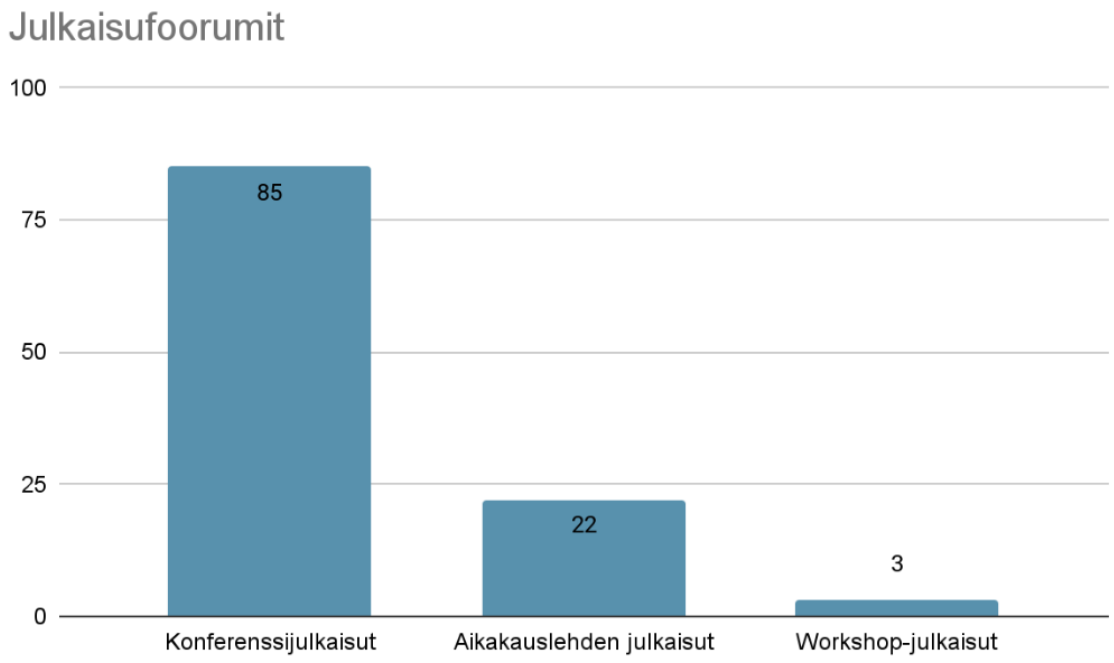
TAULUKKO 3 Julkaisufoorumit

Foorumin nimi	Tyyppi	lkm.
ACM International Conference on Management of Data (SIGMOD/PODS)	Konferenssi	16
Conference on Empirical Methods in Natural Language Processing (EMNLP)	Konferenssi	14
Association for Computational Linguistics (ACL)	Konferenssi	12
International Joint Conference on Natural Language Processing (IJCNLP)*8	Konferenssi	8
Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT)	Konferenssi	5
Very Large Data Base Endowment (VLDB Endowment)	Aikakauslehti	5
Findings of the Association for Computational Linguistics: EMNLP	Aikakauslehti	3
IEEE International Conference on Data Engineering (ICDE)	Konferenssi	3
International Conference on Computational Linguistics (COLING)	Konferenssi	3
IEEE Access	Aikakauslehti	2
International Conference for Convergence in Technology (I2CT)	Konferenssi	2
International Journal of Reasoning-based Intelligent Systems (IJRIS)	Aikakauslehti	2

ACM Conference on Hypertext and Social Media	Konferenssi	1
ACM India Joint International Conference on Data Science & Management of Data (ACM IKDD CODS & COMAD)	Konferenssi	1
ACM/SIGAPP Symposium on Applied Computing (SAC)	Konferenssi	1
Computational Linguistics	Aikakauslehti	1
Conference on Neural Information Processing Systems (NeurIPS)	Konferenssi	1
Data in Brief	Aikakauslehti	1
E3S Web of Conferences	Konferenssi	1
IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)	Konferenssi	1
IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)	Konferenssi	1
IEEE India Council International Conference (INDICON)	Konferenssi	1
IEEE International Workshop on Metrology for Industry 4.0 & IoT (MetroInd4.0&IoT)	Workshop	1
IEEE Transactions on Industrial Informatics (TII)	Aikakauslehti	1
IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)	Konferenssi	1
International Conference for Advancement in Technology (ICONAT)	Konferenssi	1
International Conference on Advanced Computing & Communication Systems (ICACCS)	Konferenssi	1
International Conference on Advances in ICT for Emerging Regions (ICter)	Konferenssi	1
International Conference on Algorithms, Computing and Artificial Intelligence (ACAI)	Konferenssi	1
International Conference on Computational Intelligence in Data Science (ICCIDS)	Konferenssi	1
International Conference on Computer and Information Technology (ICCIT)	Konferenssi	1
International Conference on Data and Software Engineering (ICoDSE)	Konferenssi	1
International Conference on Electrical Information and Communication Technology (EICT)	Konferenssi	1
International Conference on Electronics, Communication and Aerospace Technology (ICECA)	Konferenssi	1
International Conference on Information Science and Education (ICISE-IE)	Konferenssi	1
International Conference on Information Technology (ICIT)	Konferenssi	1
International Conference on Intelligent Computing in Data Sciences (ICDS)	Konferenssi	1
International Conference on Intelligent User Interfaces (IUI)	Konferenssi	1
International Conference on Language Resources and Evaluation (LREC)	Konferenssi	1
International Conference on Learning Representations (ICLR)	Konferenssi	1
International Conference on Networking, Information Systems & Security (NISS)	Konferenssi	1
International Conference on Recent Advances in Systems Science and Engineering (RASSE)	Konferenssi	1
International Joint Conference on Artificial Intelligence (IJCAI)	Konferenssi	1
International Journal for Research in Applied Science and Engineering Technology (IJRASET)	Aikakauslehti	1
International Multidisciplinary Information Technology and Engineering Conference (IMITEC)	Konferenssi	1
International Workshop on the Web and Databases	Workshop	1
IOP Conference Series: Materials Science and Engineering	Konferenssi	1

IT Professional	Aikakauslehti	1
ITM Web of Conferences	Konferenssi	1
Journal of King Saud University - Computer and Information Sciences (JKSUCI)	Aikakauslehti	1
Knowledge Representation & Reasoning Meets Machine Learning Workshop (KR2ML)	Workshop	1
Knowledge-Based Systems	Aikakauslehti	1
Neurocomputing	Aikakauslehti	1
The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS Archives)	Aikakauslehti	1
The International Journal on Very Large Data Bases (VLDB)	Aikakauslehti	1
Yhteensä		118

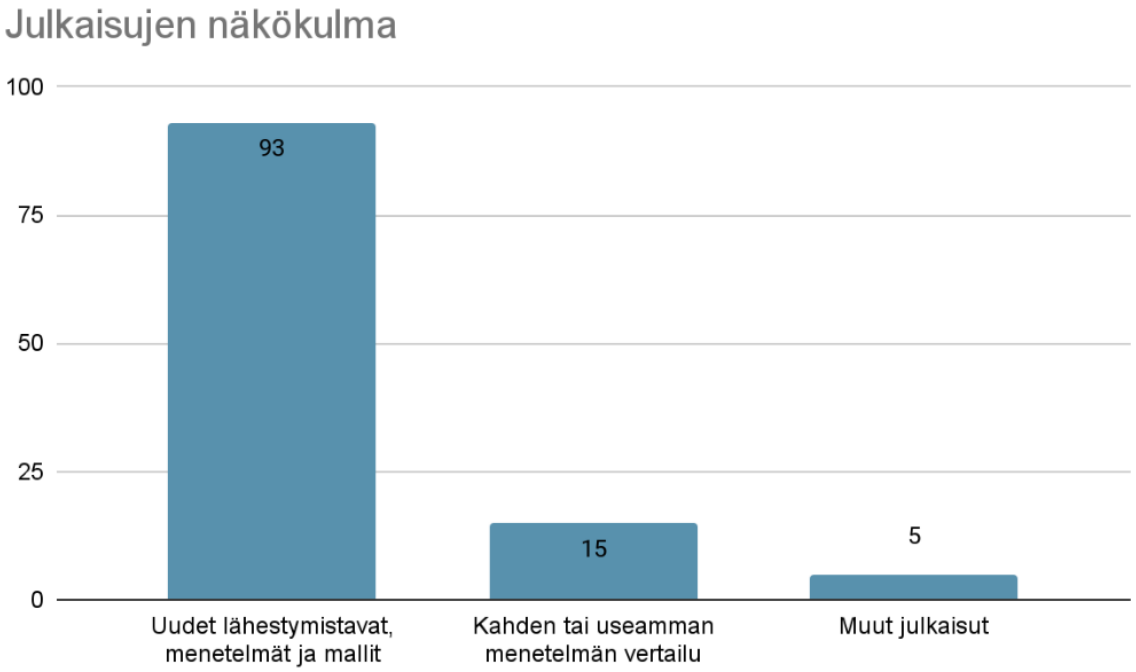
Tutkimukseen valikoituneista julkaisuista selkeästi suurin osa oli konferenssijulkaisuja; 85 oli konferenssijulkaisuja, 3 konferenssissa julkaistua workshop -tuotosta ja 22 aikakauslehden julkaisemaa julkaisua. Tutkimuksen artikkeleista yli kolmannes kuului kolmen eniten julkaisseen foorumin joukkoon. Nämä olivat kaikki konferenssijulkaisuja. Findings of the Association for Computational Linguistics: EMNLP eivät itse luokittele itseään konferenssiksi eikä aikakauslehdeksi. Tutkimuksen tekijän näkökulmasta nämä julkaisut sopivat kolmesta julkaisufoorumista parhaiten aikakauslehdeksi. Tutkielmaan valikoitui yhteensä 55 eri julkaisufoorumia. Näistä 12 julkaistiin useampi kuin yksi artikkeli. Julkaisufoorumien jakauma tulee ilmi alla esitetystä kuviosta (kuvio 3).



KUVIO 3 Julkaisufoorumit

4.3 Julkaisujen näkökulmat

Julkaisujen näkökulmat jaoteltiin kolmeen eri luokkaan. Luokista ensimmäinen on uudet lähestymistavat, menetelmät ja mallit. Tähän luokkaan luokiteltiin kaikki ne julkaisut, joiden pääasiallinen kontribuutio on uuden lähestymistavan, menetelmän, mallin tai uuden luonnollisen kielen käyttöliittymän kehittäminen tietokannoille. Näitä julkaisuja oli 93 kappaletta, joka vastaa ~85 % julkaisuista. Alta löytyvässä kuviossa (kuvio 4) on esitettynä julkaisuluokat ja niiden määrät.



KUVIO 4 Julkaisujen näkökulmat

Toiseksi eniten tutkimuksia jakautui luokkaan kahden tai useamman menetelmän vertailu. Tähän luokkaan sisällytettiin kaikki tutkimukset, joiden tarkoituksena oli vertailla kahta tai useampaa aiemmin julkaistua menetelmää tai luonnollisen kielen käyttöliittymää. Näitä tutkimuksia oli 15, joka vastaa ~14% tutkimuksista.

Viimeiseen luokkaan, muut, lueteltiin sellaiset julkaisut, joiden pääasiallinen kontribuutio oli jokin muu kuin uuden menetelmän esittely tai aiempien menetelmien vertailu keskenään. Tähän luokkaan lukeutui yhteensä 5 julkaisua. Muihin julkaisuihin luetelluista artikkeleista yksi artikkeli esitteli uuden mallin vertailutyökalusta, jonka avulla pystytään vertailemaan ja analysoimaan luonnollisen kielen tietokantajärjestelmiä. Toisessa artikkelissa tutkittiin SQL-kyselyn jakamista pienempiin osiin empiirisen tutkimuksen avulla. Kolmannessa julkaisussa esiteltiin uusi malli oppimiselle SQL-NLP:n kautta. Neljäs artikkeli keskittyi SQL-NLP:n taustatietoon ja tulevien tutkimusten kehitykseen. Lisäksi tämä artikkeli esitteli uuden mallin, joten artikkeli luokiteltiin myös ensimmäiseen luokkaan. Viides artikkeli analysoi WikiSQL:ää, suurta tietoaineistoa, jossa on

julkaistuna kymmeniä tuhansia SQL-kielen ja luonnollisen kielen kysymyspareja. Lisäksi artikkelissa julkaistiin oma ehdotus järjestelmästä WikiSQL:än käyttämiseen, jonka myötä viideskin artikkeli luokiteltiin myös ensimmäiseen luokkaan.

Yhteensä kolme julkaisua luokiteltiin useampaan kuin yhteen luokkaan, sillä julkaisujen kontribuutio sopi useampaan luokkaan. Kolmas useampaan luokkaan kuuluvista julkaisuista oli artikkeli, joka kuului ensimmäiseen ja toiseen luokkaan. Artikkelissa vertailtiin aiemmin julkaistuja menetelmiä ja luotiin oma malli luonnollisen kielen kääntämisestä SQL-kieleksi.

5 POHDINTA

Tuloksista käy ilmi, että tutkimusta SQL-NLP:stä on tehty vuosien 2018-2022 välillä runsaasti. Vaikka rajattujen julkaisuvuosien kahden viimeisimmän vuoden aikana julkaisumäärät ovat laskeneet, voidaan silti todeta aiheen olevan vieläkin relevantti. Kuten tuloksista selvisi, suurin osa uusista tutkimusta käsittelee uusien menetelmien ja lähestymistapojen esittämistä. Tämä kertoo siitä, että tietokannat ja datan käyttö lisääntyvät koko ajan erilaisissa organisaatioissa, sillä organisaatioille sopivia käyttöliittymiä luonnollisen kielen kääntämiseksi SQL-kielelle kehitetään. Ylipäätään luonnollisen kielen kehittäminen ja käyttö erilaisten tietojärjestelmien yhteydessä on kasvamaan päin. Lisäksi viime vuosien aikana on julkaistu kaikkien käytössä olevia tekoälyä hyödyntäviä chat-botteja, joiden avulla pystytään hankkimaan tietoa käyttämällä luonnollista kieltä. Tutkimusaineiston julkaisufoorumit ovat painottuneet konferensseihin, mutta myös aikakauslehtien julkaisut ovat merkittävä osa SQL-NLP:n tutkimuskenttää.

Vaikka uusia lähestymistapoja ja menetelmiä on kehitetty luonnollisen kielen kääntämiseksi SQL-kielelle, niin toisia menetelmiä keskenään vertailevia artikkeleita on julkaistu suhteellisen vähän. Monessa aineiston artikkelissa keskityttiin oman uuden menetelmän positiivisiin puoliin ja korostettiin, kuinka oman menetelmän käytöllä pystytään saada parempia tuloksia kuin toisilla menetelmillä. Kim, So, Han ja Lee (2020) toteavat tutkimuksessaan, että tutkimuksia on luonnollisen kielen kääntämisestä SQL-kieleksi tehty paljon, mutta perusteellista käsitystä tekniikoiden todellisesta tietoa niiden hyödyntämisestä ja käytännön tilanteista ei ole tarpeeksi. Keskeisimmäksi vaikeudeksi vertailuun he nostavat tutkimusten erilaiset tietokokonaisuudet. Monessa eri tutkimuksessa keskitytään tiettyyn sektoriin, eikä samaa menetelmää pystytä hyödyntämään erilaisten tietokokonaisuuksien kanssa. Tämän myötä olisi tärkeää pyrkiä keksimään yhteneväiset mittarit menetelmien mittaamiselle (Kim ym., 2020). Samaan johtopäätökseen tulevat Gkini, Belmpas, Koutrika ja Ioannidis tutkimuksessaan (2021). He toteavat, että olemassa olevat menetelmien arvioinnit käyttävät erilaisia tietojoukkoja ja kyselyitä, jonka myötä menetelmien tutkimisessa on myös erilaiset arviointikriteerit ja -tavoitteet. Heidän tutkimuksessaan, An In-Depth Benchmarking of Text-to-SQL Systems, he pyrkivät vastaamaan tähän esitettyyn

ongelmaan. Lisäksi heidän tavoitteenaan on paljastaa uusia avoimia haasteita luonnollisen kielen kääntämisessä SQL-kyselyiksi (Gkini ym., 2021). Tämä onkin kaikista laajin useaa SQL-NLP-menetelmää vertaileva tutkimus. Tutkimuksessa vertaillaan yhteensä kahdeksan eri järjestelmän tehokkuutta. Yhteenvedossa tutkijat mainitsevat, että tulevaisuudessa tavoitteena on laajentaa testiä käsittelemään useampia kyselytyyppejä (Gkini ym., 2021).

Muita edellä mainitun kaltaisia relevantteja laajoja vertailututkimuksia ei löytynyt tämän tutkimuksen aineistosta. Julkaisussaan Gkini ja muut (2021) esittävät toiveen siitä, että heidän tutkimuksensa loisi pohjaa systemaattisemmalle luonnollisen kielen kääntämisen arvioinnille. Lisäksi he toivovat, että heidän työnsä inspiroi muita tutkimuksia ja järjestelmiä (Gkini ym., 2021).

6 YHTEENVETO

Tämän kandidaatin tutkielman tarkoituksena oli selvittää, millaista SQL-NLP-tutkimusta on tehty viimeisien vuosien aikana. Tutkimusmenetelmänä käytettiin systemaattista kirjallisuuskartoitusta, jonka avulla saatiin selville käsiteltävän aiheen merkittävimmät julkaisut rajatuin kriteerein. Systemaattisessa kirjallisuuskartoituksessa saatu aineisto luokitellaan ja lajitellaan, jonka jälkeen saadaan selkeä kuva aiemman tutkimuksen määrästä ja tulokulmista. Menetelmän tarkoituksena on kartoittaa aiempaa tutkimusta ja saada objektiivinen kuva valitun aiheen tutkimuskentästä. Tutkimusmenetelmään ei sisälly systemaattisen kirjallisuuskatsauksen tapaan tavoitetta syvällisestä ymmärtämisestä koskien valittua aihetta, vaan systemaattisella kirjallisuuskartoituksella tähdätään korkeamman tason luokitteluun.

Aineiston hankinnassa hyödynnettiin neljää tietokantaa, joissa suoritettiin tietokantahaut. Tietokantahaut rajattiin vuosien 2018-2022 välille, jolloin hakutuloksina saatiin yhteensä 829 julkaisua. Julkaisuja käsiteltiin otsikkotason, tiivistelmän ja koko tekstin perusteella. Tietokantahauista soveltuviin julkaisuihin hyödynnettiin taaksepäin suuntautuvaa lumipallomenetelmää. Toisin sanoen tietokantahakujen soveltuvien artikkelien lähdeluettelot käytiin läpi, jonka avulla saatiin uusia sopivia julkaisuja tutkimukseen. Tämän jälkeen kaikkia julkaisuja analysointiin kriittisesti pohjaten valittuihin aineiston valintakriteereihin. Lopullinen tutkimusaineisto koostui 110 tieteellisestä vertaisarvioidusta artikkelista, jotka käsittelevät luonnollisen kielen hyödyntämistä SQL-kielessä. Saatu aineisto luokiteltiin tutkimuskysymysten pohjalta julkaisuvuoden, julkaisufoorumin ja tutkimuksen näkökulman perusteella.

Tutkimuksen tuloksista selvisi, että aiheesta tehtyä tutkimusta on tehty runsaasti viimeisen viiden vuoden aikana. Eniten tieteellisiä julkaisuja aiheesta oli julkaistu vuonna 2020. Valikoituneista julkaisuista suurin osa oli konferenssijulkaisuja. Vain vajaa neljännes oli aikakauslehden julkaisuja tai workshop-julkaisuja. Kolmas saatu tulos liittyi tutkimusten näkökulmiin. Saadusta aineistosta selvisi, että suurin osa uudesta tutkimuksesta keskittyy uusien menetelmien kehittämiseen luonnollisen kielen kääntämiseksi SQL-kielelle. Tutkimuksia on julkaistu myös näkökulmasta, jossa vertaillaan kehiteltyjä menetelmiä. Laajoja

vertailevia tutkimuksia on kuitenkin tehty suhteellisen vähän. Samaan johtopäätökseen vertailevien tutkimuksien harvuudesta päädyttiin myös usean muun tutkimuksen tuloksista, joita tähän tutkimukseen valikoitui.

Tutkielmassa suurin riski validiteetin ja yleistettävyyden kannalta on aineistohaun valinta vain yhden tutkijan toimesta. Useamman kuin yhden tutkijan tapauksessa aineiston valinnasta pystytään keskustelemaan ja rajatapauksia pystytään arvioimaan paremmin. Toinen suuri riski on aiheen laajuus kandidaatin tutkielmassa. Jälkeenpäin kriittisesti ajateltuna aihetta olisi voitu rajata pienemmäksi, jolloin kohtuullisen työmäärän puitteissa olisi pystytty suorittamaan eteenpäin suuntautuvaa lumipallomenetelmää, jossa uudempia tuotoksia olisi etsitty seuraamalla sitaatteja. Tämä olisi ollut tärkeää etenkin tämän tutkimuksen tapauksessa, koska tietokantahauissa käytetty hakulauseke oli yksinkertainen eikä sitä lähdetty laventamaan liiallisen työmäärän vuoksi. Tämä saattoi vaikuttaa tutkimuksen julkaisuihin niin, että kaikkia relevantteja vuosien 2021 ja 2022 julkaisuja ei välttämättä saatu sisällytettyä tutkimukseen . Tämä voi heikentää tutkimusten tulosten yleistettävyyttä.

Aiempaa tutkimusta luonnollisen kielen kääntämisestä SQL-kieleksi systemaattisen kirjallisuuskartoituksen osalta ei ollut tehty ennen tätä tutkimusta. Tämä systemaattinen kirjallisuuskartoitus toimii hyvänä pohjana tuleville tutkimuksille, koska tutkimuksessa on koottuna yhteen kaikki merkittävät tutkimukset aiheesta.

Monessa tutkimuksessa, jossa on esitelty uusi kääntämismenetelmä tai -tekniikka, on saatu hyviä tuloksia. Kuitenkaan kriittisiä, montaa tekijää huomioonottavia vertailevia tutkimuksia, ei ole näistä tekniikoista viime vuosina julkaistu tarpeeksi. Tulosten perusteella lisätutkimusta siis tarvitaan erilaisten menetelmien vertailusta luonnollisen kielen kääntämisestä SQL-kieleksi. Monessa vertailevassa tutkimuksessa vertailu on rajattu vain tiettyyn toimintaan luonnollisen kielen käyttämisessä SQL-kielessä ja relaatiotietokannoissa. Tämän myötä toisena mielenkiintoisena jatkotutkimusaiheena olisi analysoida ja koota yhteen tutkimuksia, joissa vertaillaan eri menetelmien toimivuutta ja tehokkuutta. Tämä voitaisiin suorittaa käyttämällä menetelmänä systemaattista kirjallisuuskartoitusta.

LÄHTEET

Budgen, D., Turner, M., Brereton, P., & Kitchenham, B. (2008). Using Mapping Studies in Software Engineering. *Proceedings of PPIG 2008*, *2*.

Deshpande, A., Kothari, D., Salvi, A., Mane, P., & Kolhe, V. (2022). Querylizer: An Interactive Platform for Database Design and Text to SQL Conversion. *2022 International Conference for Advancement in Technology (ICONAT)*, 1–6. https://doi.org/10.1109/ICONAT53423.2022.9725828

Gkini, O., Belmpas, T., Koutrika, G., & Ioannidis, Y. (2021). An In-Depth Benchmarking of Text-to-SQL Systems. *Proceedings of the 2021 International Conference on Management of Data*, 632–644. https://doi.org/10.1145/3448016.3452836

Karimi, S., Rasel, A. A., & Abdullah, M. S. (2022). Non-English Natural Language Interface to Databases: A Systematic Review. *2022 IEEE 13th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 0391–0397. https://doi.org/10.1109/IEMCON56893.2022.9946569

Katsogiannis-Meimarakis, G., & Koutrika, G. (2021). A Deep Dive into Deep Learning Approaches for Text-to-SQL Systems. *Proceedings of the 2021 International Conference on Management of Data*, 2846–2851. https://doi.org/10.1145/3448016.3457543

Kim, H., So, B.-H., Han, W.-S., & Lee, H. (2020). Natural language to SQL: Where are we today? *Proceedings of the VLDB Endowment*, *13*(10), 1737–1750. https://doi.org/10.14778/3401960.3401970

Kitchenham, B., Budgen, D., & Brereton, P. (2010). The value of mapping studies—A participant-observer case study. *Proceedings of the 14th International Conference on Evaluation and Assessment in Software Engineering*.

Petersen, K., Vakkalanka, S., & Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, *64*, 1–18. https://doi.org/10.1016/j.infsof.2015.03.007

Reshma, E. U., & Remya, P. C. (2017). A review of different approaches in natural language interfaces to databases. *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, 801–804. https://doi.org/10.1109/ISS1.2017.8389287

Sanyal, H., Shukla, S., & Agrawal, R. (2021). Natural Language Processing Technique for Generation of SQL Queries Dynamically. *2021 6th International Conference for Convergence in Technology (I2CT)*, 1–6. https://doi.org/10.1109/I2CT51068.2021.9418091

Shah, D., Das, A., Shahane, A., Parikh, D., & Bari, P. (2021). SpeakQL Natural Language to SQL. *ITM Web of Conferences*, *40*, 03018. https://doi.org/10.1051/itmconf/20214003018

Shah, V., Li, S., Kumar, A., & Saul, L. (2020). SpeakQL: Towards Speech-driven Multimodal Querying of Structured Data. *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 2363–2374. https://doi.org/10.1145/3318464.3389777

Silva, Y. N., Almeida, I., & Queiroz, M. (2016). SQL: From Traditional Databases to Big Data. *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, 413–418. https://doi.org/10.1145/2839509.2844560

Taipalus, T. (2023). Systematic Mapping Study in Information Systems Research. *Journal of the Midwest Association for Information Systems (JMWAIS)*, *2023*(1). https://doi.org/10.17705/3jmwa.000079

Uma, M., Sneha, V., Sneha, G., Bhuvana, J., & Bharathi, B. (2019). Formation of SQL from Natural Language Query using NLP. *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, 1–5. https://doi.org/10.1109/ICCIDS.2019.8862080

Wilson, D.-M., Martin, A. M., & Gilbert, J. E. (2010). 'How may I help you'-spoken queries for technical assistance. *Proceedings of the 48th Annual Southeast Regional Conference*, 1–6. https://doi.org/10.1145/1900008.1900068

Wohlin, C., Kalinowski, M., Romero Felizardo, K., & Mendes, E. (2022). Successful combination of database search and snowballing for identification of primary studies in systematic literature studies. *Information and Software Technology*, *147*, 106908. https://doi.org/10.1016/j.infsof.2022.106908

LIITE 1 TAULUKKO TUTKIMUSAINEISTOSTA

Taulukko tutkimusaineistoksi valikoituneista julkaisuista.
JF = Julkaisufoorumi, 1 = Konferenssijulkaisu, 2 = Aikakauslehden julkaisu
3 = Workshop-julkaisu konferenssissa

NK = Näkökulma, A = Uudet lähestymistavat, menetelmät ja mallit, B = Kahden tai useamman menetelmän vertailu C = Muut

Julkaisu	JF	NK
Affolter, K., Stockinger, K., & Bernstein, A. (2019). A Comparative Survey of Recent Natural Language Interfaces for Databases. *The VLDB Journal*, *28*(5), 793–819. https://doi.org/10.1007/s00778-019-00567-8	2	B
Ahkouk, K., & Machkour, M. (2019). Human Language Question To SQL Query Using Deep Learning. *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*, 1–6. https://doi.org/10.1109/ICDS47004.2019.8942342	1	B
Ahkouk, K., & Machkour, M. (2020). Towards an interface for translating natural language questions to SQL: A conceptual framework from a systematic review. *International Journal of Reasoning-based Intelligent Systems*, *12*, 264–275. https://doi.org/10.1504/IJRIS.2020.111786	2	AB
Ahkouk, K., Machkour, M., Khadija, M., & Mama, R. (2020). SEQ2SEQ VS SKETCH FILLING STRUCTURE FOR NATURAL LANGUAGE TO SQL TRANSLATION. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLIV-4/W3-2020*. https://doi.org/10.5194/isprs-archives-XLIV-4-W3-2020-7-2020	2	B
Ahkouk, K., Mustapha, M., Khadija, M., & Rachid, M. (2021). A review of the Text to SQL Frameworks. *Proceedings of the 4th International Conference on Networking, Information Systems & Security*, 1–6. https://doi.org/10.1145/3454127.3457619	1	A
Ahmed, M., & Uddin, M. N. (2020). Cyber Attack Detection Method Based on NLP and Ensemble Learning Approach. *2020 23rd*	1	A

International Conference on Computer and Information Technology (ICCIT), 1–6. https://doi.org/10.1109/ICCIT51783.2020.9392682		
Al-Muhammed, M. J., & Lonsdale, D. W. (2022). Ontology-aware dynamically adaptable free-form natural language agent interface for querying databases. *Knowledge-Based Systems*, *239*, 108012. https://doi.org/10.1016/j.knosys.2021.108012	2	A
Ananthanarayanan, R., Lohia, P. K., & Bedathur, S. (2018). DataVizard: Recommending Visual Presentations for Structured Data. *Proceedings of the 21st International Workshop on the Web and Databases*, 1–6. https://doi.org/10.1145/3201463.3201465	3	A
Anisyah, A., Widagdo, T. E., & Nur Azizah, F. (2019). Natural Language Interface to Database (NLIDB) for Decision Support Queries. *2019 International Conference on Data and Software Engineering (ICoDSE)*, 1–6. https://doi.org/10.1109/ICoDSE48700.2019.9092769	1	A
Baik, C., Jagadish, H. V., & Li, Y. (2019). Bridging the Semantic Gap with SQL Query Logs in Natural Language Interfaces to Databases. *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, 374–385. https://doi.org/10.1109/ICDE.2019.00041	1	A
Baik, C., Jin, Z., Cafarella, M., & Jagadish, H. V. (2020). Duoquest: A Dual-Specification System for Expressive SQL Queries. *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 2319–2329. https://doi.org/10.1145/3318464.3389776	1	A
Barsha, M. K., Azharul Hasan, K. M., & Ara, I. (2021). Natural Language Interface to Database by Regular Expression Generation. *2021 5th International Conference on Electrical Information and Communication Technology (EICT)*, 1–6. https://doi.org/10.1109/EICT54103.2021.9733592	1	A
Basik, F., Hättasch, B., Ilkhechi, A., Usta, A., Ramaswamy, S., Utama, P., Weir, N., Binnig, C., & Cetintemel, U. (2018). DBPal: A Learned NL-Interface for Databases. *Proceedings of the 2018 International Conference on Management of Data*, 1765–1768. https://doi.org/10.1145/3183713.3193562	1	A
Belmpas, T., Gkini, O., & Koutrika, G. (2020). Analysis of Database Search Systems with THOR. *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 2681–2684. https://doi.org/10.1145/3318464.3384679	1	C
Bogin, B., Berant, J., & Gardner, M. (2019). Representing Schema Structure with Graph Neural Networks for Text-to-SQL Parsing. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4560–4565. https://doi.org/10.18653/v1/P19-1448	1	A
Bogin, B., Gardner, M., & Berant, J. (2019). Global Reasoning over Database Structures for Text-to-SQL Parsing. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3650–3655. https://doi.org/10.18653/v1/D19-1378	1	A

Brunner, U., & Stockinger, K. (2021). ValueNet: A Natural Language-to-SQL System that Learns from Database Information. *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, 2177–2182. https://doi.org/10.1109/ICDE51399.2021.00220	1	A
Cai, R., Xu, B., Zhang, Z., Yang, X., Li, Z., & Liang, Z. (2018). An encoder-decoder framework translating natural language to database queries. *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 3977–3983.	1	A
Cai, R., Yuan, J., Xu, B., & Hao, Z. (2021). *SADGA: Structure-Aware Dual Graph Aggregation Network for Text-to-SQL* (arXiv:2111.00653). arXiv. https://doi.org/10.48550/arXiv.2111.00653	1	A
Cai, Y., & Wan, X. (2020). IGSQL: Database Schema Interaction Graph Based Neural Model for Context-Dependent Text-to-SQL Generation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6903–6912. https://doi.org/10.18653/v1/2020.emnlp-main.560	1	A
Cao, R., Chen, L., Chen, Z., Zhao, Y., Zhu, S., & Yu, K. (2021). LGESQL: Line Graph Enhanced Text-to-SQL Model with Mixed Local and Non-Local Relations. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2541–2555. https://doi.org/10.18653/v1/2021.acl-long.198	1	A
Chen, S., San, A., Liu, X., & Ji, Y. (2020). A Tale of Two Linkings: Dynamically Gating between Schema Linking and Structural Linking for Text-to-SQL Parsing. *Proceedings of the 28th International Conference on Computational Linguistics*, 2900–2912. https://doi.org/10.18653/v1/2020.coling-main.260	1	A
Chen, Z., Chen, L., Zhao, Y., Cao, R., Xu, Z., Zhu, S., & Yu, K. (2021). ShadowGNN: Graph Projection Neural Network for Text-to-SQL Parser. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5567–5577. https://doi.org/10.18653/v1/2021.naacl-main.441	1	A
Choi, D., Shin, M. C., Kim, E., & Shin, D. R. (2021). RYANSQL: Recursively Applying Sketch-based Slot Fillings for Complex Text-to-SQL in Cross-Domain Databases. *Computational Linguistics*, *47*(2), 309–332. https://doi.org/10.1162/coli_a_00403	2	A
Das, A., & Balabantaray, R. C. (2019). MyNLIDB: A Natural Language Interface to Database. *2019 International Conference on Information Technology (ICIT)*, 234–238. https://doi.org/10.1109/ICIT48102.2019.00048	1	A
Deng, X., Awadallah, A. H., Meek, C., Polozov, O., Sun, H., & Richardson, M. (2021). Structure-Grounded Pretraining for Text-to-SQL. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*	1	A

Technologies, 1337–1350. https://doi.org/10.18653/v1/2021.naacl-main.105		
Deshpande, A., Kothari, D., Salvi, A., Mane, P., & Kolhe, V. (2022). Querylizer: An Interactive Platform for Database Design and Text to SQL Conversion. *2022 International Conference for Advancement in Technology (ICONAT)*, 1–6. https://doi.org/10.1109/ICONAT53423.2022.9725828	1	A
Dong, L., & Lapata, M. (2018). Coarse-to-Fine Decoding for Neural Semantic Parsing. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 731–742. https://doi.org/10.18653/v1/P18-1068	1	A
Dong, Z., Sun, S., Liu, H., Lou, J.-G., & Zhang, D. (2019). Data-Anonymous Encoding for Text-to-SQL Generation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5405–5414. https://doi.org/10.18653/v1/D19-1543	1	A
Ferreira, S., Leitão, G., Silva, I., Martins, A., & Ferrari, P. (2020). Evaluating Human-Machine Translation with Attention Mechanisms for Industry 4.0 Environment SQL-Based Systems. *2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT*, 229–234. https://doi.org/10.1109/MetroInd4.0IoT48571.2020.9138181	3	A
Finegan-Dollak, C., Kummerfeld, J. K., Zhang, L., Ramanathan, K., Sadasivam, S., Zhang, R., & Radev, D. (2018). Improving Text-to-SQL Evaluation Methodology. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 351–360. https://doi.org/10.18653/v1/P18-1033	1	B
Gan, Y., Chen, X., Xie, J., Purver, M., Woodward, J. R., Drake, J., & Zhang, Q. (2021). Natural SQL: Making SQL Easier to Infer from Natural Language Specifications. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2030–2042. https://doi.org/10.18653/v1/2021.findings-emnlp.174	2	A
Gkini, O., Belmpas, T., Koutrika, G., & Ioannidis, Y. (2021). An In-Depth Benchmarking of Text-to-SQL Systems. *Proceedings of the 2021 International Conference on Management of Data*, 632–644. https://doi.org/10.1145/3448016.3452836	1	B
Godinez, J. E., & Jamil, H. M. (2019). Meet cyrus: The query by voice mobile assistant for the tutoring and formative assessment of SQL learners. *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, 2461–2468. https://doi.org/10.1145/3297280.3297523	1	A
Gogoi, B., Ahmed, T., & Dutta, A. (2021). Defending against SQL Injection Attacks in Web Applications using Machine Learning and Natural Language Processing. *2021 IEEE 18th India Council International Conference (INDICON)*, 1–6. https://doi.org/10.1109/INDICON52576.2021.9691740	1	A

Guo, A., Zhao, X., & Ma, W. (2021). ER-SQL: Learning enhanced representation for Text-to-SQL using table contents. *Neurocomputing*, *465*, 359–370. https://doi.org/10.1016/j.neucom.2021.08.134	2	A
Guo, J., Zhan, Z., Gao, Y., Xiao, Y., Lou, J.-G., Liu, T., & Zhang, D. (2019). Towards Complex Text-to-SQL in Cross-Domain Database with Intermediate Representation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4524–4535. https://doi.org/10.18653/v1/P19-1444	1	A
Gur, I., Yavuz, S., Su, Y., & Yan, X. (2018). DialSQL: Dialogue Based Structured Query Generation. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1339–1349. https://doi.org/10.18653/v1/P18-1124	1	A
Hains, G. J. D. R., Khmelevsky, Y., & Tachon, T. (2019). From natural language to graph queries. *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*, 1–4. https://doi.org/10.1109/CCECE.2019.8861892	1	A
Hosu, I. A., Iacob, R. C. A., Brad, F., Ruseti, S., & Rebedea, T. (2018). Natural Language Interface for Databases Using a Dual-Encoder Model. *Proceedings of the 27th International Conference on Computational Linguistics*, 514–524. https://aclanthology.org/C18-1043	1	A
Huang, P.-S., Wang, C., Singh, R., Yih, W., & He, X. (2018). Natural Language to Structured Query Generation via Meta-Learning. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 732–738. https://doi.org/10.18653/v1/N18-2115	1	C
Hwang, W., Yim, J., Park, S., & Seo, M. (2019). *A Comprehensive Exploration on WikiSQL with Table-Aware Word Contextualization* (arXiv:1902.01069). arXiv. https://doi.org/10.48550/arXiv.1902.01069	3	A
Iacob, R. C. A., Brad, F., Apostol, E.-S., Truică, C.-O., Hosu, I. A., & Rebedea, T. (2020). Neural Approaches for Natural Language Interfaces to Databases: A Survey. *Proceedings of the 28th International Conference on Computational Linguistics*, 381–395. https://doi.org/10.18653/v1/2020.coling-main.34	1	B
Jammi, M., Sen, J., Mittal, A., Verma, S., Pahuja, V., Ananthanarayanan, R., Lohia, P., Karanam, H., Saha, D., & Sankaranarayanan, K. (2018). Tooling framework for instantiating natural language querying system. *Proceedings of the VLDB Endowment*, *11*(12), 2014–2017. https://doi.org/10.14778/3229863.3236248	2	A
Joshi, S. R., Venkatesh, B., Thomas, D., Jiao, Y., & Roy, S. (2020). A Natural Language and Interactive End-to-End Querying and Reporting System. *Proceedings of the 2021 International Conference on Management of Data*, 261–267. https://doi.org/10.1145/3371158.3371198	1	A
Kaoshik, R., Patil, R., R, P., Agarawal, S., Jain, N., & Singh, M. (2021). ACL-SQL: Generating SQL Queries from Natural Language.	1	A

Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD), 423. https://doi.org/10.1145/3430984.3431046		
Karimi, S., Rasel, A. A., & Abdullah, M. S. (2022b). Non-English Natural Language Interface to Databases: A Systematic Review. *2022 IEEE 13th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 0391–0397. https://doi.org/10.1109/IEMCON56893.2022.9946569	1	B
Kate, A., Kamble, S., Bodkhe, A., & Joshi, M. (2018). Conversion of Natural Language Query to SQL Query. *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 488–491. https://doi.org/10.1109/ICECA.2018.8474639	1	A
Katsogiannis-Meimarakis, G., & Koutrika, G. (2021). A Deep Dive into Deep Learning Approaches for Text-to-SQL Systems. *Proceedings of the 2021 International Conference on Management of Data*, 2846–2851. https://doi.org/10.1145/3448016.3457543	1	B
Kedwan, F. (2022). NLQ into SQL translation using computational linguistics. *Journal of King Saud University - Computer and Information Sciences*, *34*(9), 6564–6582. https://doi.org/10.1016/j.jksuci.2022.03.010	2	A
Khadija, M., & Machkour, M. (2021). The history and recent advances of Natural Language Interfaces for Databases Querying. *E3S Web of Conferences*, *229*, 01039. https://doi.org/10.1051/e3sconf/202122901039	1	B
Kim, H., So, B.-H., Han, W.-S., & Lee, H. (2020). Natural language to SQL: Where are we today? *Proceedings of the VLDB Endowment*, *13*(10), 1737–1750. https://doi.org/10.14778/3401960.3401970	2	B
Koutti, L., Machkour, M., & Bais, H. (2018). An Arabic natural language interface for querying relational databases based on natural language processing and graph theory methods. *International Journal of Reasoning-based Intelligent Systems*, *10*, 155. https://doi.org/10.1504/IJRIS.2018.10013299	2	A
Lakhani, S., Yadav, A., & Singh, V. (2022). Detecting SQL Injection Attack using Natural Language Processing. *2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, 1–5. https://doi.org/10.1109/UPCON56432.2022.9986458	1	A
Lee, D. (2019). Clause-Wise and Recursive Decoding for Complex and Cross-Domain Text-to-SQL Generation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6045–6051. https://doi.org/10.18653/v1/D19-1624	1	A
Lei, W., Wang, W., Ma, Z., Gan, T., Lu, W., Kan, M.-Y., & Chua, T.-S. (2020). Re-examining the Role of Schema Linking in Text-to-SQL. *Proceedings of the 2020 Conference on Empirical Methods in Natural*	1	AC

Language Processing (EMNLP), 6943–6954. https://doi.org/10.18653/v1/2020.emnlp-main.564		
Li, Q., Li, L., Li, Q., & Zhong, J. (2020). A Comprehensive Exploration on Spider with Fuzzy Decision Text-to-SQL Model. *IEEE Transactions on Industrial Informatics*, *16*(4), 2542–2550. https://doi.org/10.1109/TII.2019.2952929	2	A
Lin, V., Socher, R., & Xiong, C. (2020). *Bridging Textual and Tabular Data for Cross-Domain Text-to-SQL Semantic Parsing*. 4870–4888. https://doi.org/10.18653/v1/2020.findings-emnlp.438	2	A
Liu, H., Fang, L., Liu, Q., Chen, B., Lou, J.-G., & Li, Z. (2019). Leveraging Adjective-Noun Phrasing Knowledge for Comparison Relation Prediction in Text-to-SQL. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3515–3520. https://doi.org/10.18653/v1/D19-1356	1	A
Liu, Q., Yang, D., Zhang, J., Guo, J., Zhou, B., & Lou, J.-G. (2021). Awakening Latent Grounding from Pretrained Language Models for Semantic Parsing. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1174–1189. https://doi.org/10.18653/v1/2021.findings-acl.100	1	A
Liu, X., He, P., Chen, W., & Gao, J. (2019). Multi-Task Deep Neural Networks for Natural Language Understanding. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4487–4496. https://doi.org/10.18653/v1/P19-1441	1	A
Long, H., & Cao, D. (2021). Bert-based Text-to-SQL Generation method with question-table content enhancement and template filling. *2021 2nd International Conference on Information Science and Education (ICISE-IE)*, 969–973. https://doi.org/10.1109/ICISE-IE53922.2021.00222	1	A
Luo, Y., Tang, N., Li, G., Chai, C., Li, W., & Qin, X. (2021). Synthesizing Natural Language to Visualization (NL2VIS) Benchmarks from NL2SQL Benchmarks. *Proceedings of the 2021 International Conference on Management of Data*, 1235–1247. https://doi.org/10.1145/3448016.3457261	1	A
Ma, J., Yan, Z., Pang, S., Zhang, Y., & Shen, J. (2020). Mention Extraction and Linking for SQL Query Generation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6936–6942. https://doi.org/10.18653/v1/2020.emnlp-main.563	1	A
Ma, P., & Wang, S. (2021). MT-teql: Evaluating and augmenting neural NLIDB on real-world linguistic and schema variations. *Proceedings of the VLDB Endowment*, *15*(3), 569–582. https://doi.org/10.14778/3494124.3494139	2	B
Narechania, A., Fourney, A., Lee, B., & Ramos, G. (2021). DIY: Assessing the Correctness of Natural Language to SQL Systems. *26th*	1	A

International Conference on Intelligent User Interfaces, 597–607. https://doi.org/10.1145/3397481.3450667		
Obaido, G., Ade-Ibijola, A., & Vadapalli, H. (2020). TalkSQL: A Tool for the Synthesis of SQL Queries from Verbal Specifications. *2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, 1–10. https://doi.org/10.1109/IMITEC50163.2020.9334088	1	A
Pal, D., Sharma, H., & Chaudhuri, K. (2021). Data Agnostic RoBERTa-based Natural Language to SQL Query Generation. *2021 6th International Conference for Convergence in Technology (I2CT)*, 1–5. https://doi.org/10.1109/I2CT51068.2021.9417888	1	A
Peduru Hewa, D. S., & Farook, C. (2021). A Sinhala Natural Language Interface for Querying Databases Using Natural Language Processing. *2021 21st International Conference on Advances in ICT for Emerging Regions (ICter)*, 213–218. https://doi.org/10.1109/ICter53630.2021.9774794	1	A
Publication, I. (2020). Natural Language Processing with some Abbreviation to SQL. *International Journal for Research in Applied Science and Engineering Technology - IJRASET*. https://www.academia.edu/42192716/Natural_Language_Processing_with_some_Abbreviation_to_SQL	2	A
Salimzadeh, S., Gadiraju, U., Hauff, C., & van Deursen, A. (2022). Exploring the Feasibility of Crowd-Powered Decomposition of Complex User Questions in Text-to-SQL Tasks. *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, 154–165. https://doi.org/10.1145/3511095.3531282	1	C
Sanyal, H., Shukla, S., & Agrawal, R. (2021). Natural Language Processing Technique for Generation of SQL Queries Dynamically. *2021 6th International Conference for Convergence in Technology (I2CT)*, 1–6. https://doi.org/10.1109/I2CT51068.2021.9418091	1	A
Scholak, T., Li, R., Bahdanau, D., de Vries, H., & Pal, C. (2021). DuoRAT: Towards Simpler Text-to-SQL Models. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1313–1321. https://doi.org/10.18653/v1/2021.naacl-main.103	1	A
Sen, J., Lei, C., Quamar, A., Özcan, F., Efthymiou, V., Dalmia, A., Stager, G., Mittal, A., Saha, D., & Sankaranarayanan, K. (2020). ATHENA++: Natural language querying for complex nested SQL queries. *Proceedings of the VLDB Endowment*, *13*(12), 2747–2759. https://doi.org/10.14778/3407790.3407858	2	A
Sen, J., Ozcan, F., Quamar, A., Stager, G., Mittal, A., Jammi, M., Lei, C., Saha, D., & Sankaranarayanan, K. (2019). Natural Language Querying of Complex Business Intelligence Queries. *Proceedings of the 2019 International Conference on Management of Data*, 1997–2000. https://doi.org/10.1145/3299869.3320248	1	A

Shah, D., Das, A., Shahane, A., Parikh, D., & Bari, P. (2021). SpeakQL Natural Language to SQL. *ITM Web of Conferences*, *40*, 03018. https://doi.org/10.1051/itmconf/20214003018	1	A
Shah, V., Li, S., Kumar, A., & Saul, L. (2020). SpeakQL: Towards Speech-driven Multimodal Querying of Structured Data. *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 2363–2374. https://doi.org/10.1145/3318464.3389777	1	A
Shah, V., Li, S., Yang, K., Kumar, A., & Saul, L. (2019). Demonstration of SpeakQL: Speech-driven Multimodal Querying of Structured Data. *Proceedings of the 2019 International Conference on Management of Data*, 2001–2004. https://doi.org/10.1145/3299869.3320224	1	A
Sheinin, V., Khorashani, E., Yeo, H., Xu, K., Vo, N. P. A., & Popescu, O. (2018). QUEST: A Natural Language Interface to Relational Databases. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. LREC 2018, Miyazaki, Japan. https://aclanthology.org/L18-1469	1	A
Shi, T., Zhao, C., Boyd-Graber, J., III, H., & Lee, L. (2020). *On the Potential of Lexico-logical Alignments for Semantic Parsing to SQL Queries*. 1849–1864. https://doi.org/10.18653/v1/2020.findings-emnlp.167	2	A
Singh, H. (2019). Interfaces to Query Relational Databases in Natural Language. *IT Professional*, *21*(1), 67–73. https://doi.org/10.1109/MITP.2018.2876983	2	B
Song, M., Zhan, Z., & E., H. (2019). Hierarchical Schema Representation for Text-to-SQL Parsing With Decomposing Decoding. *IEEE Access*, *7*, 103706–103715. https://doi.org/10.1109/ACCESS.2019.2931464	2	A
Song, Y., Wong, R. C.-W., Zhao, X., & Jiang, D. (2022). VoiceQuerySystem: A Voice-driven Database Querying System Using Natural Language Questions. *Proceedings of the 2022 International Conference on Management of Data*, 2385–2388. https://doi.org/10.1145/3514221.3520158	1	A
Sowah, E., & Xu, J. (2018). Edgebase: A Cooperative Query Answering Database System With A Natural Language Interface. *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence*, 1–8. https://doi.org/10.1145/3302425.3302482	1	A
Sun, Y., Tang, D., Duan, N., Ji, J., Cao, G., Feng, X., Qin, B., Liu, T., & Zhou, M. (2018). Semantic Parsing with Syntax- and Table-Aware SQL Generation. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 361–372. https://doi.org/10.18653/v1/P18-1034	1	A
Tomova, M. T., Hofmann, M., & Mäder, P. (2022). SEOSS-Queries—A software engineering dataset for text-to-SQL and question answering tasks. *Data in Brief*, *42*, 108211. https://doi.org/10.1016/j.dib.2022.108211	2	A

Uma, M., Sneha, V., Sneha, G., Bhuvana, J., & Bharathi, B. (2019b). Formation of SQL from Natural Language Query using NLP. *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, 1–5. https://doi.org/10.1109/ICCIDS.2019.8862080	1	A
Usta, A., Karakayali, A., & Ulusoy, Ö. (2021). DBTagger: Multi-task learning for keyword mapping in NLIDBs using Bi-directional recurrent neural networks. *Proceedings of the VLDB Endowment*, *14*(5), 813–821. https://doi.org/10.14778/3446095.3446103	2	A
Wang, B., Shin, R., Liu, X., Polozov, O., & Richardson, M. (2020). RAT-SQL: Relation-Aware Schema Encoding and Linking for Text-to-SQL Parsers. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7567–7578. https://doi.org/10.18653/v1/2020.acl-main.677	1	A
Wang, W., Bhowmick, S. S., Li, H., Joty, S., Liu, S., & Chen, P. (2021). Towards Enhancing Database Education: Natural Language Generation Meets Query Execution Plans. *Proceedings of the 2021 International Conference on Management of Data*, 1933–1945. https://doi.org/10.1145/3448016.3452822	1	A
Wang, W., Tian, Y., Wang, H., & Ku, W.-S. (2020). A Natural Language Interface for Database: Achieving Transfer-learnability Using Adversarial Method for Question Understanding. *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, 97–108. https://doi.org/10.1109/ICDE48307.2020.00016	1	A
Wei, Z., Trummer, I., & Anderson, C. (2021). Demonstrating Robust Voice Querying with MUVE: Optimally Visualizing Results of Phonetically Similar Queries. *Proceedings of the 2021 International Conference on Management of Data*, 2798–2802. https://doi.org/10.1145/3448016.3452753	1	A
Weir, N., & Utama, P. (2019). Bootstrapping an End-to-End Natural Language Interface for Databases. *Proceedings of the 2019 International Conference on Management of Data*, 1862–1864. https://doi.org/10.1145/3299869.3300105	1	A
Weir, N., Utama, P., Galakatos, A., Crotty, A., Ilkhechi, A., Ramaswamy, S., Bhushan, R., Geisler, N., Hättasch, B., Eger, S., Cetintemel, U., & Binnig, C. (2020). DBPal: A Fully Pluggable NL2SQL Training Pipeline. *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 2347–2361. https://doi.org/10.1145/3318464.3380589	1	A
Wong, A., Joiner, D., Chiu, C., Elsayed, M., Pereira, K., Khmelevsky, Y., & Mahony, J. (2021). A Survey of Natural Language Processing Implementation for Data Query Systems. *2021 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE)*, 1–8. https://doi.org/10.1109/RASSE53195.2021.9686815	1	B
Wudaru, V., Koditala, N., Reddy, A., & Mamidi, R. (2019). Question Answering on Structured Data using NLIDB Approach. *2019 5th*	1	A

International Conference on Advanced Computing & Communication Systems (ICACCS), 1–4. https://doi.org/10.1109/ICACCS.2019.8728487		
Xu, B., Cai, R., Zhang, Z., Yang, X., Hao, Z., Li, Z., & Liang, Z. (2019). NADAQ: Natural Language Database Querying Based on Deep Learning. *IEEE Access*, *7*, 35012–35017. https://doi.org/10.1109/ACCESS.2019.2904720	2	A
Yao, Z., Su, Y., Sun, H., & Yih, W. (2019). Model-based Interactive Semantic Parsing: A Unified Framework and A Text-to-SQL Case Study. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5447–5458. https://doi.org/10.18653/v1/D19-1547	1	A
Yavuz, S., Gur, I., Su, Y., & Yan, X. (2018). What It Takes to Achieve 100% Condition Accuracy on WikiSQL. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1702–1711. https://doi.org/10.18653/v1/D18-1197	1	AC
Yu, T., Li, Z., Zhang, Z., Zhang, R., & Radev, D. (2018). TypeSQL: Knowledge-Based Type-Aware Neural Text-to-SQL Generation. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 588–594. https://doi.org/10.18653/v1/N18-2093	1	A
Yu, T., Wu, C.-S., Lin, X. V., Wang, B., Tan, Y. C., Yang, X., Radev, D., Socher, R., & Xiong, C. (2021, toukokuuta 28). *GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing*. https://doi.org/10.48550/arXiv.2009.13845	1	A
Yu, T., Yasunaga, M., Yang, K., Zhang, R., Wang, D., Li, Z., & Radev, D. (2018). SyntaxSQLNet: Syntax Tree Networks for Complex and Cross-Domain Text-to-SQL Task. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1653–1663. https://doi.org/10.18653/v1/D18-1193	1	A
Yu, T., Zhang, R., Er, H., Li, S., Xue, E., Pang, B., Lin, X. V., Tan, Y. C., Shi, T., Li, Z., Jiang, Y., Yasunaga, M., Shim, S., Chen, T., Fabbri, A., Li, Z., Chen, L., Zhang, Y., Dixit, S., … Radev, D. (2019). CoSQL: A Conversational Text-to-SQL Challenge Towards Cross-Domain Natural Language Interfaces to Databases. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1962–1979. https://doi.org/10.18653/v1/D19-1204	1	A
Yu, T., Zhang, R., Yang, K., Yasunaga, M., Wang, D., Li, Z., Ma, J., Li, I., Yao, Q., Roman, S., Zhang, Z., & Radev, D. (2018). *Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task*. 3911–3921. https://doi.org/10.18653/v1/D18-1425	1	A

Yu, T., Zhang, R., Yasunaga, M., Tan, Y. C., Lin, X. V., Li, S., Er, H., Li, I., Pang, B., Chen, T., Ji, E., Dixit, S., Proctor, D., Shim, S., Kraft, J., Zhang, V., Xiong, C., Socher, R., & Radev, D. (2019). SParC: Cross-Domain Semantic Parsing in Context. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4511–4523. https://doi.org/10.18653/v1/P19-1443	1	A
Zeng, J., Lin, X. V., Hoi, S. C. H., Socher, R., Xiong, C., Lyu, M., & King, I. (2020). Photon: A Robust Cross-Domain Text-to-SQL System. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 204–214. https://doi.org/10.18653/v1/2020.acl-demos.24	1	A
Zhang, R., Yu, T., Er, H., Shim, S., Xue, E., Lin, X. V., Shi, T., Xiong, C., Socher, R., & Radev, D. (2019). Editing-Based SQL Query Generation for Cross-Domain Context-Dependent Questions. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5338–5349. https://doi.org/10.18653/v1/D19-1537	1	A
Zhekova, M., & Totkov, G. (2019). Conceptual Frame Model For The Presentation Of The Concepts And Rules In Natural Language Interface For Database. *IOP Conference Series: Materials Science and Engineering*, *618*(1), 012035. https://doi.org/10.1088/1757-899X/618/1/012035	1	A
Zhong, R., Yu, T., & Klein, D. (2020). Semantic Evaluation for Text-to-SQL with Distilled Test Suites. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 396–411. https://doi.org/10.18653/v1/2020.emnlp-main.29	1	A
Őzcan, F., Quamar, A., Sen, J., Lei, C., & Efthymiou, V. (2020). State of the Art and Open Challenges in Natural Language Interfaces to Data. *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 2629–2636. https://doi.org/10.1145/3318464.3383128	1	B